International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)
Volume 4, Issue 4, 2017, PP 53-63
ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online)
DOI: http://dx.doi.org/10.20431/2349-4859.0404007
www.arcjournals.org

Understanding the Dimensions of Data on Domain Basis

Dr. Srinatha Karur[1], Prof. M. V. Ramana Murthy[2]

Department of Informatics, Wollega Univeristy, Nekemtee, Ethiopia

Department of Mathematics, Osmania University, Hyderabad, India

Corresponding Author:** ***Dr. Srinatha Karur*, Department of Informatics, Wollega Univeristy, Nekemtee, Ethiopia*

Abstract: *In this paper the authors are attempt to try estimate the dimensions of the available data. Even though the real time data nature is more costly and time consuming and confidential .Sometimes it is necessary to understand the data perfectly for provisional decisions. Most of the real time data is always influenced by multiple dimensions which are independent on each other. In this paper the authors try to establish the balance point for Normal Distribution. and other Distributions.*

Keywords: *Lattice, Cube theory, Distributions and Data mining tools*

1. Introduction

It is very challenging task to define the scope the real time data for any domain or problem. That scope is completely depend upon not only on the context of availability of data and also dimensions of the data and levels of the data. Using Mathematical concepts and other relevant tools we can understand the real nature of data. Especially for unstructured decisions it is very difficult to understand the exact nature of data[1]. Sometimes it is showing multiple solutions for the single value of data as per the standard definition of Operations Research. In Data mining it is more complicated to understand exact nature of data due to its hidden pattern. While data transformation there is variation in data values. The authors published paper on outlier's formation on the basis of "Data mining techniques" for real time data. In this paper the authors pointed out the various situations or forms of data for cluster analysis or cluster preparation with real time data[2]. The real time applications are interpret with Business Intelligence and other Management domains. So it is necessary to study the various factors which influence the data for decision making. It is possible to find the nature of data up to some extension with the help of server level tools or client level tools. But server level implementation gives more depth on understanding and different phases of data. For this purpose the authors prefer either MS-SQL server or Oracle server Data mining. Along with this the Oracle SQL or MS-SQL also preferable to understand the exact nature of given problem. The most popular software tools are also available for Data distribution analysis for understand the real depth of data and its multi dimensions nature. The computer science community generally using R language or Tanagra for Statistical Analysis.

2. Methods/Literature Review

As per standard definition of Decision theory it is very clear about the different types of decisions which are available as part of system. But when domain application changes its interpretation also very difficult. Since there are lot of variety of situations are available [3].

Customer Relationship Management: A Database Approach offers the promise of maximized profits for today's highly competitive businesses. This innovative book provides readers with the tools and techniques to effectively use CRM. It emphasizes the utilization of database marketing in order to build strong and profitable customer relationships. Kumar first describes how to implement database marketing and then looks at recent advances in CRM applications. Critical marketing issues like optimum resource allocation, purchase sequence, and the link between acquisition, retentions, and profitability are also examined on the basis of empirical findings. Some authors or real time data experts express their view as “the data dimensions are nothing but the customer expectation from quality of the product or service nature of company and also the highly influenced by market trends [4].

If we take or consider any organization in the business world generally it has three layer or pyramid approach and influenced by marketing trends and nature and buying capacity of the customer. The authors think that implementation of the data for Management Information System is out of scope of topic as MIS have its own scope and independent on Business Intelligence. But it gives CRM to partial guidelines for data analysis to take appropriate decisions. Along with the MIS the authors discussed about the different components of CRM for its Data base preparation and process of customer life cycle [5].

In terms of Data mining the curve fitting method is extremely useful to find out the relation between two variables where as in business world it is not giving the exact logical meaning between two variables. Very standard algorithms and procedures are available from research community and official web sites. The readers or research scholars or students can follow any University level prescribed text books. Generally research community has interest on two dimension scope instead of n-dimension scope. For multi dimension analysis the readers can follow the MATLAB software for engineering applications. The readers can get more information n-dimension from MATLAB official web site [6].

2.1. Methods Discussed

For real understand of data for n-dimension it is very good to understand the Vector space model mathematically which mainly deals about the n-dimension space. In mathematics, the dimension of a vector space V is the cardinality (i.e. the number of vectors) of a basis of V over its base field. For every vector space there exists a basis, and all bases of a vector space have equal cardinality; as a result, the dimension of a vector space is uniquely defined [7]. It means that the context should be unique and free from ambiguity nature. In real time this type of decisions are possible with structured decisions in terms of MIS and in terms of Operation Research it is called Decision under certainty. For more details on Vector space model refer the respected websites and refer standard University level books.

Decision theory in economics, psychology, philosophy, mathematics, and statistics is concerned with identifying the values, uncertainties and other issues relevant in a given decision, its rationality, and the resulting optimal decision. It is closely related to the field of game theory as to interactions of agents with at least partially conflicting interests whose decisions affect each other [8].

In Computational theory there are different classes are available and these classes are applicable as per given problem. Each and every problems restricted with some rules or dimensions for achieve the required goal or answer. There are lot of problems are available which are not yet solved by computer due to volatile dimensions of given problem [9].

In Data warehouse and Data mining technologies the dimensions of data is explained on the basis of cube theory. Where cube consists of three dimensions and each and every dimension is treated as array in programming languages. The standard algorithms are available for cube theory and academic point of view. But practically or in real world some deviations are available. The deviations of real world approach and pure algorithms for cube theory. Very good standard official documents are available for cube implementation in terms of Oracle and Microsoft SQL-Servers. The details of server level cube implementation are out of scope of paper. Lattice theory is also very helpful for discuss the nature of cubes in data warehouse. In Data mining we can have techniques to reduce the dimensions of the data[10]. Along with this the authors are also discussed the different strategies of cube computing techniques [10,pg:192-194].

The students or readers can get very good results of cube theory with the help of Oracle-SQL and MS-SQL also. In the business world. Along with this SQL tools also available for implement the data cubes with star schema. The following is the best example from Wikipedia of star schema. It is the simplest and more popular on the data cube construction.

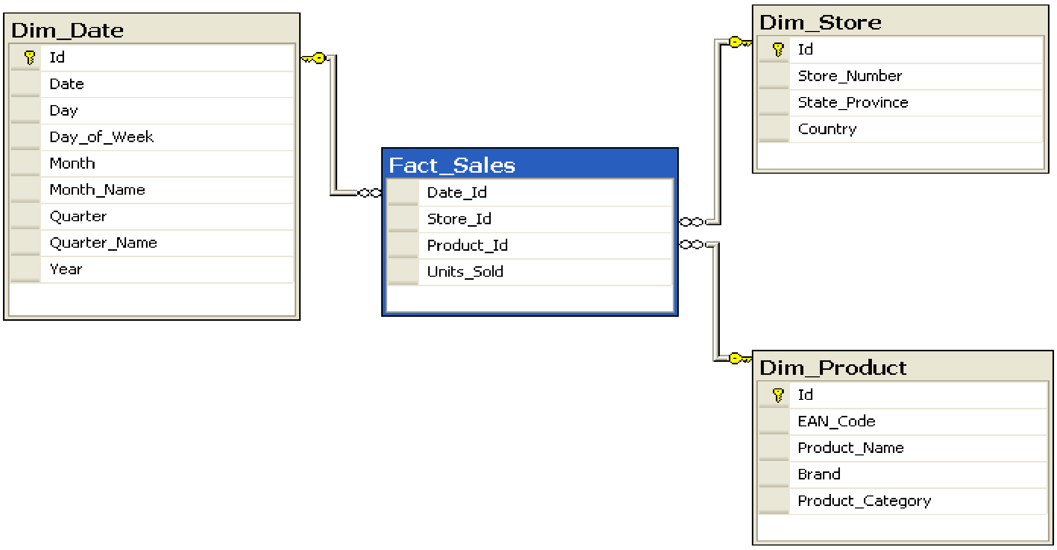
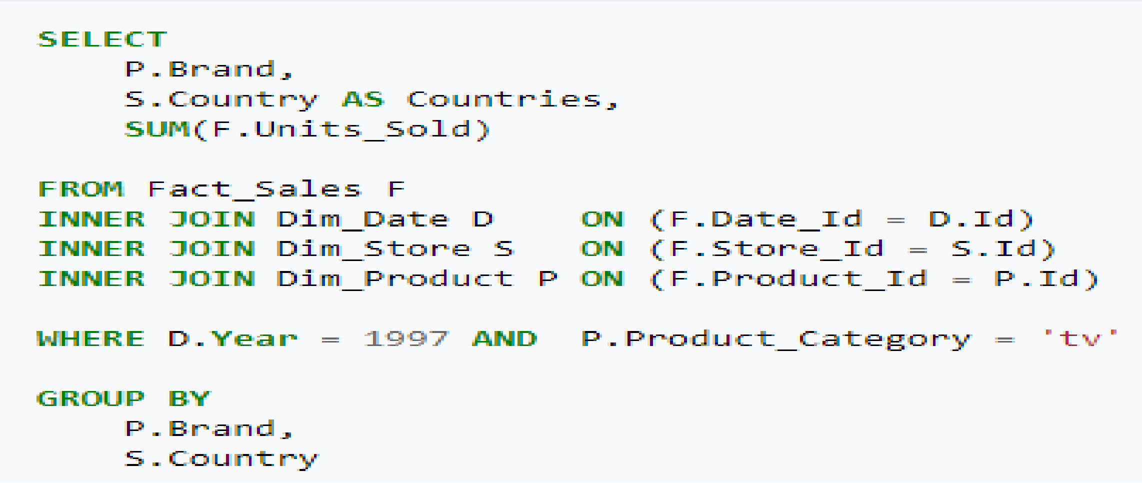


Figure1. *Shows star schema for data cube(3 dimensions)*



```
SELECT
    P.Brand,
    S.Country AS Countries,
    SUM(F.Units_Sold)

FROM Fact_Sales F
INNER JOIN Dim_Date D    ON (F.Date_Id = D.Id)
INNER JOIN Dim_Store S   ON (F.Store_Id = S.Id)
INNER JOIN Dim_Product P ON (F.Product_Id = P.Id)

WHERE D.Year = 1997 AND  P.Product_Category = 'tv'

GROUP BY
    P.Brand,
    S.Country
```

Figure2. *SQL for cube of 3 dimensions*

The recent technologies are implement this cube on the basis of star schema technology. Mainly data cubes for Business Intelligence and customer expected values. The data cubes are also useful for increase the SQL performance. Star schema is only for construct the Data warehouses with some strategies. The required SQL is always depend upon expected output. Cube theory is implemented always on the basis of GROUP BY clause family irrespective of technologies. Only aggregative functions are applied on GROUP BY family clauses. Detailed information is available in Oracle and MS-Server official web sites.

Mathematically the Lattice theory is highly useful for dealing of dealing the dimensions of data and its relations and subsets. The detailed mathematical properties of Lattice theory is out of scope of this paper. Let us assume {a,b,c} is a small given set then the lattice can be constructed for given set is as shown in the below figure 3[11].

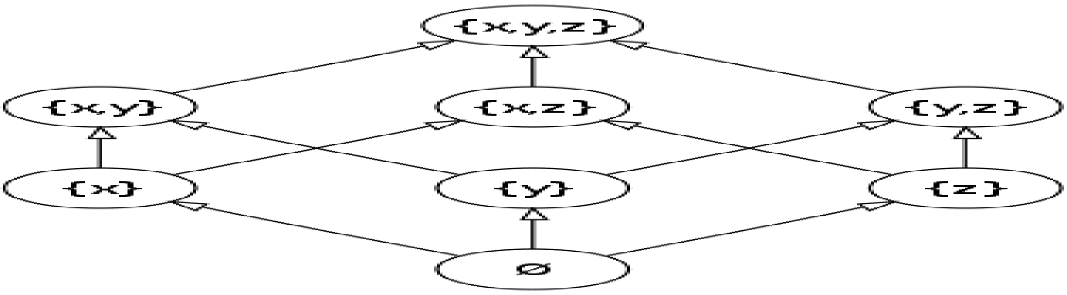


Figure3. *Shows lattice for given set*

Automated drawing tool and details are available in[12]. These tools are very particular about three dimensions only and higher orders are not possible and they are required more high level Graph theory knowledge. The coverage of Graph theory concepts is out of scope of this paper’s aim.

The goal of GA is to find a solution to a complex optimization problem, which optimal or near-optimal. GA searches for better perform candidates, where performance can be measured in terms of objective ”fitness” functions. The objective is to choose the proportion of financial assets to hold in a portfolio such that risk is minimized given the constraint of achieving a specified level of return. Fitness function does as a metrics to measure scheduler performance for each chromosome in the problem [13]. The genetic algorithm works on optimization criteria and chooses the best individuals and these best individuals are always from best fitness function only.

FCA method was devised in the early 1980s by R. Wille, a German mathematician and professor. He used the philosophical interpretation of the concept as a unit of thought comprising of a set of objects and a set of their shared attributes. Theoretical foundations of FCA are built on applied lattice theory and set theory[14].

The lattice package, written by Deepayan Sarkar, attempts to improve on base R graphics by providing better defaults and the ability to easily display multivariate relationships. In particular, the package supports the creation of trellis graphs - graphs that display a variable or the relationship between variables, conditioned on one or more other variables[15].The authors used the four dimensions from the real world problem and they are mileage, weight, number of gears, number of cylinders. Then the graph is as follows (sample graph).

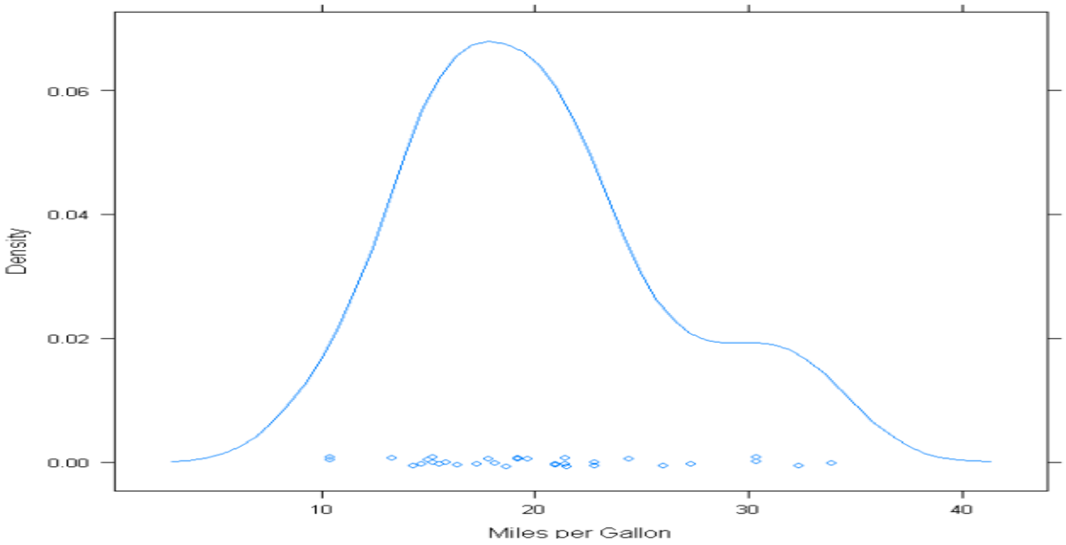


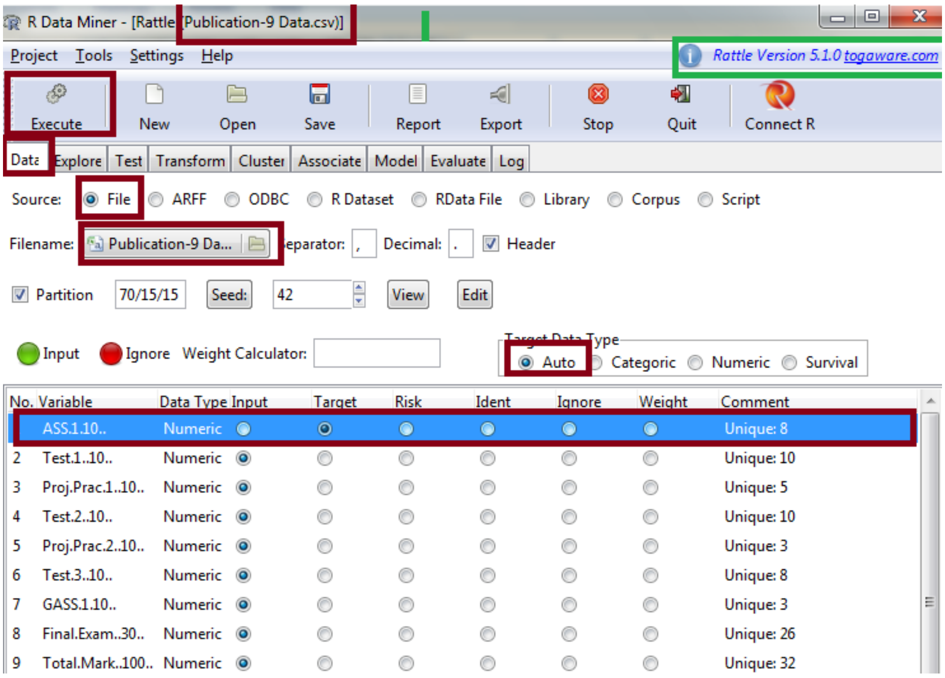
Figure4. *Shows sample graph of Lattice with 4 dimensions*

The authors are advised to students and research scholars is that we can apply the lattice data to probability distributions and the get the graphical curve results. The implementation details are out of scope and users can refer R documentation and other Statistical resources. For more information refer[16] for data analysis. Especially Lattice graphics curves refer [17]. This is available only available as online resource. The authors advised the audience to check the official web site of R for more Data analysis for different dimensions with graphics curves [18]. The authors found that the more implementations are available in [19] on Lattice graphics.

In the Business Intelligence the four dimensions are very important to chive the objective function of the company or real world problem. The competitors, customer’s nature, market value and leading professional constitute the Business Intelligence system. These four dimensions are sometimes static and dynamic some times. It is very difficult to construct lattice with this type of dimensions and maximum 2 n subsets are available as per lattice theory. But practically or in real time the subsets also leads the other subsets and hence contradiction takes place. Along with these four factors the author in his book[20] pointed out another three factors as dimensions which are highly contradict to Data mining.

3. Results and Discussion

In this section the authors have a plan to conduct the experiment on the basis of Data mining and pure Statistical methods. First the authors loaded the real data into the “R-Rattle”. By default it takes first assignment as target variable. But the readers can change the target variable as any column value and understand the interpretation as per applied conditions. The columns are available as shown in the figure 5.


R Data Miner - [Rattle Publication-9 Data.csv)]
Project Tools Settings Help
Rattle Version 5.1.0 togaware.com
Execute
New
Open
Save
Report
Export
Stop
Quit
Connect R
Data Explore Test Transform Cluster Associate Model Evaluate Log

Source: ● File ○ ARFF ○ ODBC ○ R Dataset ○ RData File ○ Library ○ Corpus ○ Script

Filename: Publication-9 Da... Separator: , Decimal: . ☑ Header

☑ Partition 70/15/15 Seed: 42 View Edit

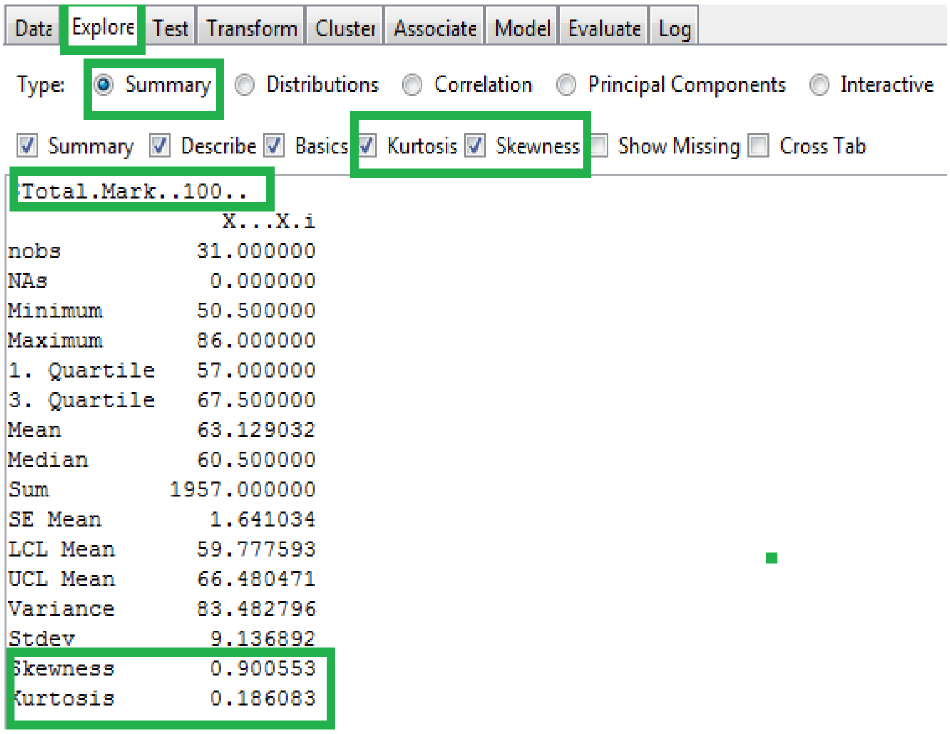
● Input ● Ignore Weight Calculator:

Target Data Type: ● Auto ○ Categoric ○ Numeric ○ Survival

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
	ASS.1.10..	Numeric	○	●	○	○	○	○	Unique: 8
2	Test.1..10..	Numeric	●	○	○	○	○	○	Unique: 10
3	Proj.Prac.1..10..	Numeric	●	○	○	○	○	○	Unique: 5
4	Test.2..10..	Numeric	●	○	○	○	○	○	Unique: 10
5	Proj.Prac.2..10..	Numeric	●	○	○	○	○	○	Unique: 3
6	Test.3..10..	Numeric	●	○	○	○	○	○	Unique: 8
7	GASS.1.10..	Numeric	●	○	○	○	○	○	Unique: 3
8	Final.Exam..30..	Numeric	●	○	○	○	○	○	Unique: 26
9	Total.Mark..100..	Numeric	●	○	○	○	○	○	Unique: 32

Figure5. *Shows the data is loaded into R-Rattle tool*

The authors have been chosen the .csv file for experiment and .arff is standard type file accepted by almost all data mining tools. The .arff file is also highly preferable standard format for experiment. From the above figure the readers can easily find out the name of the file, target data type, target column, version of the software and detailed menu ribbon. The Execute button is useful to execute the data successfully. Figure-5 shows loading and execution of data is successfully completed.


Data Explore Test Transform Cluster Associate Model Evaluate Log

Type: ● Summary ○ Distributions ○ Correlation ○ Principal Components ○ Interactive

☑ Summary ☑ Describe ☑ Basics ☑ Kurtosis ☑ Skewness □ Show Missing □ Cross Tab

```
Total.Mark..100..
                  X...X.i
nobs            31.000000
NAs              0.000000
Minimum         50.500000
Maximum         86.000000
1. Quartile     57.000000
3. Quartile     67.500000
Mean            63.129032
Median          60.500000
Sum           1957.000000
SE Mean          1.641034
LCL Mean        59.777593
UCL Mean        66.480471
Variance        83.482796
Stdev            9.136892
Skewness         0.900553
Kurtosis         0.186083
```

Figure6. *Shows Kurtosis and Skewness test*

Figure 6 shows all Statistical details where Mean, Standard Deviation, Skewness and Kurtosis is available. If Skewness is>0 means the data has positive Skewness (bended towards positive direction). Whereas standard normal curve has zero Skewness. The Kurtosis deals about the shape of the curve. The more details are available in [21]. In this paper the author's got the acceptable values of Research community opinion of the values of Kurtosis between 0 and 3 depend upon the dimensions of the data and type of software and different formulas of Kurtosis. More information is available in [22] for the accepted values of Kurtosis for given data [23]

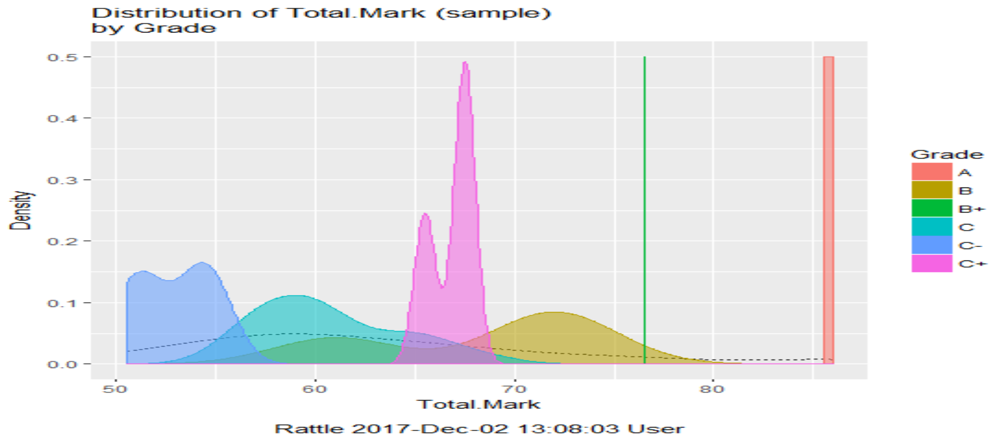


Figure7. *Shows histograms for given data with R-Rattle*

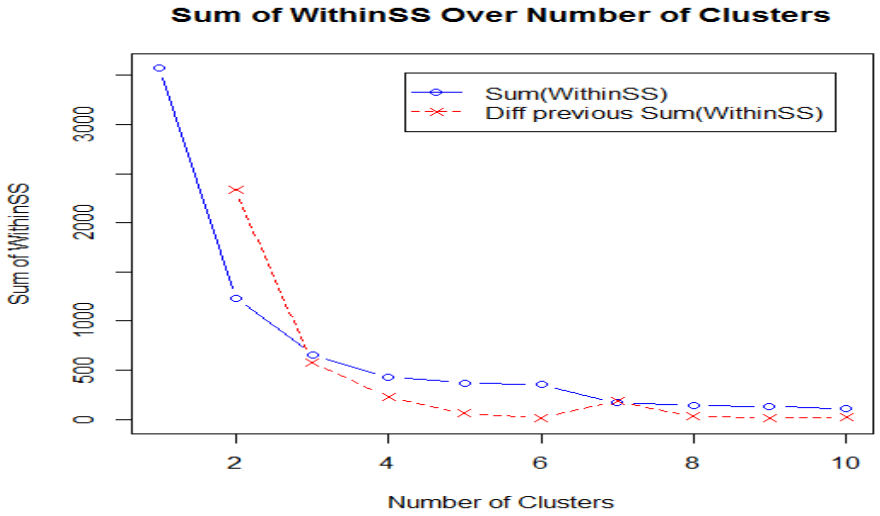


Figure8. *Shows for Iterate the ten clusters with sum of squares(K-means)*

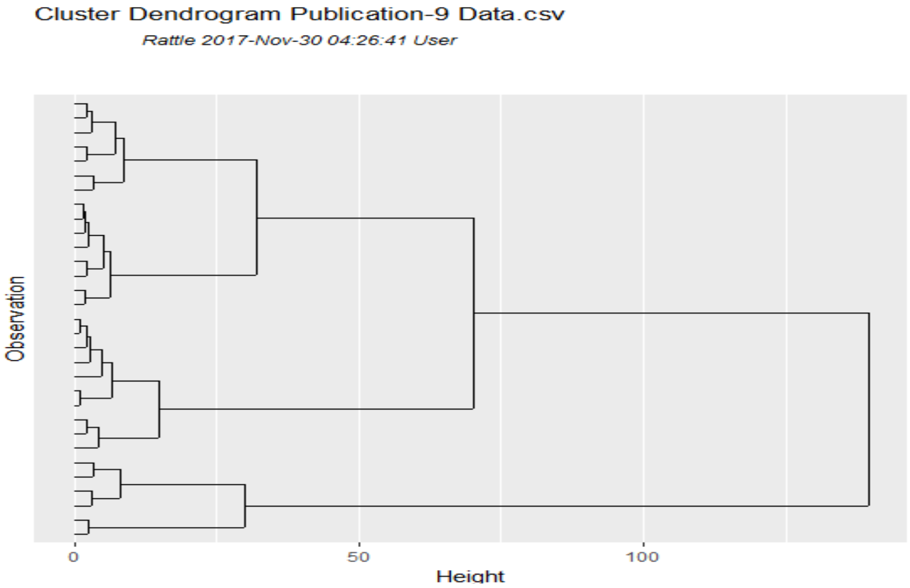


Figure9. *Shows dendogram for given data (Hierarchical clustering)*

The authors first examine for unsupervised methods(K-Means, Hierarchical clusters)We have iterated over cluster sizes from 3 to 10 clusters. The plot displays the 'sum(withinss)' for each clustering and the change in this value from the previous clustering. Minimum 3 clusters are necessary else no effect of iteration of clusters with seed value 100. The K-means clusters method is very popular in Unsupervised methods and it is very flexible. The Sum of Square is very useful to fit the data for required function. The Sum of Square method is also useful in Neural Networks construction. The authors think as it is pure mathematical model and out of scope .

The authors observed the dendogram with the following settings. The cluster method is “ ward”, the distance method is “Euclidian” and objects are 31. The authors found disciminent tab in R and it is discriminate coordinates in both K-means and Hierarchical clustering are both same i.e. 68.07% point of variability which is shown in below figure-9.

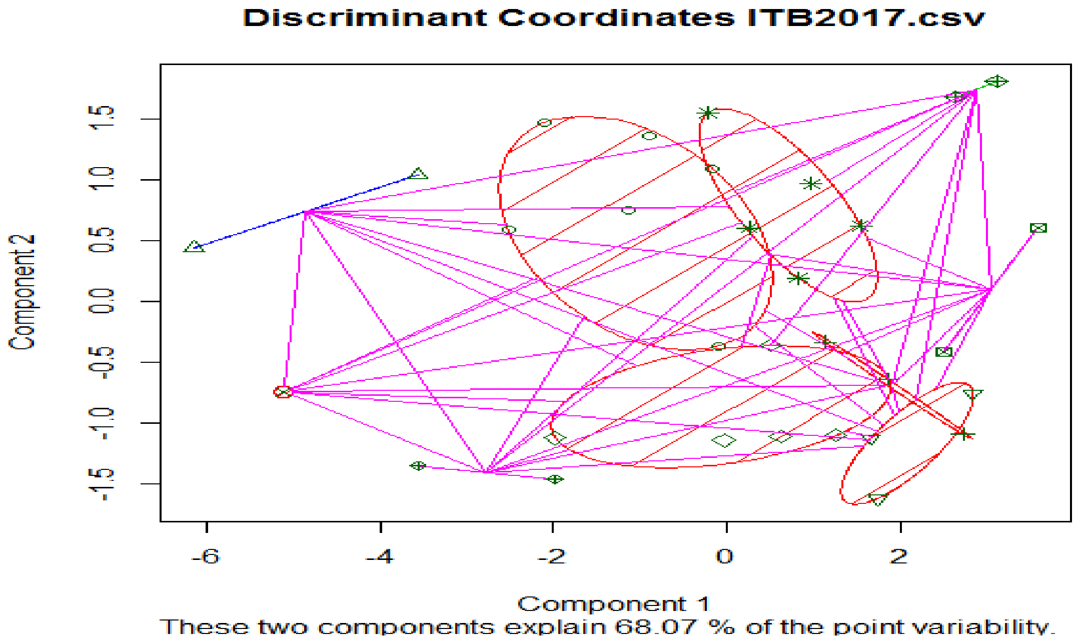
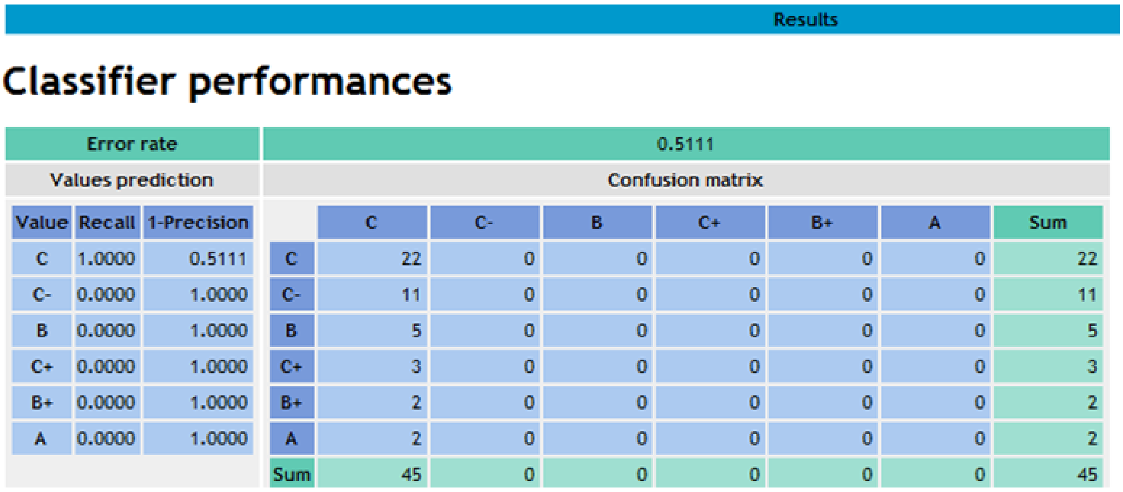


Figure10. *Shows discriminate coordinates with 68.07% point variability.*



Results

Classifier performances

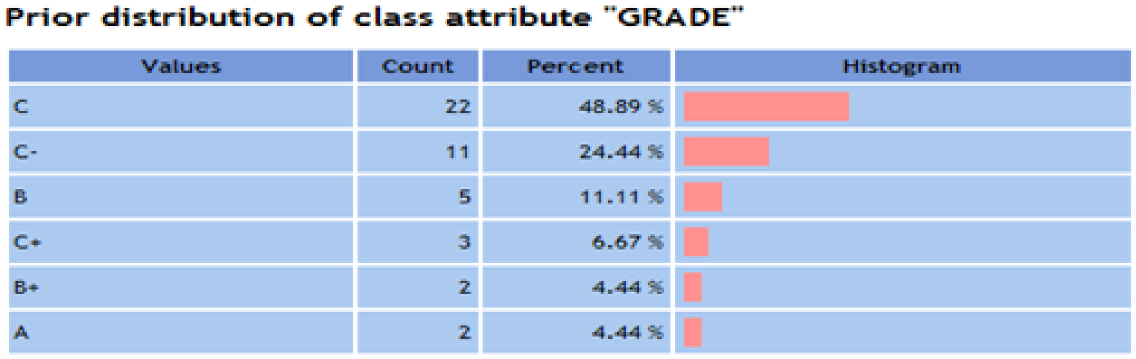
Error rate	0.5111
Values prediction	Confusion matrix

Value	Recall	1-Precision
C	1.0000	0.5111
C-	0.0000	1.0000
B	0.0000	1.0000
C+	0.0000	1.0000
B+	0.0000	1.0000
A	0.0000	1.0000

	C	C-	B	C+	B+	A	Sum
C	22	0	0	0	0	0	22
C-	11	0	0	0	0	0	11
B	5	0	0	0	0	0	5
C+	3	0	0	0	0	0	3
B+	2	0	0	0	0	0	2
A	2	0	0	0	0	0	2
Sum	45	0	0	0	0	0	45

Figure11. *Shows Confusion matrix for Naïve-Bayes method*

The authors next applied the Naïve-Bayes method(Supervised method with TANAGRA) for the given data. The confusion matrix and other methods are also available for evaluation of classification methods. They are recall and precision and they are available under prediction column. The details of Recall and Precision are out of scope of this paper and confusion matrix only examined as one of the matrices of the Naïve-Bayes method. The classifier performance has 0.5% error in their training and testing of data set. The readers can get more information on Tanagra software from[25]. More technical information about Tanagra and UNIX flavor operating systems is available in [26].



Prior distribution of class attribute "GRADE"

Values	Count	Percent	Histogram
C	22	48.89 %	
C-	11	24.44 %	
B	5	11.11 %	
C+	3	6.67 %	
B+	2	4.44 %	
A	2	4.44 %	

Figure12. *Histogram for Grade column with Naïve-Bayes method*

The authors observed the histograms generated by Naïve-Bayes method and it has very friendly information all histograms are available on the basis of Grade class. The “count” value gives the number of students belongs to under that grade. From figure-11 it is found that there are 6 types of grades are available and total 45 students are successfully enrolled for course and completed their course. None of students are get “F” grade.

Finally the authors applied the data for Statistical analysis with R and MS-Excel. The “Normal Distribution and the Histograms” are obtained are as follows (figure13,14,&15).

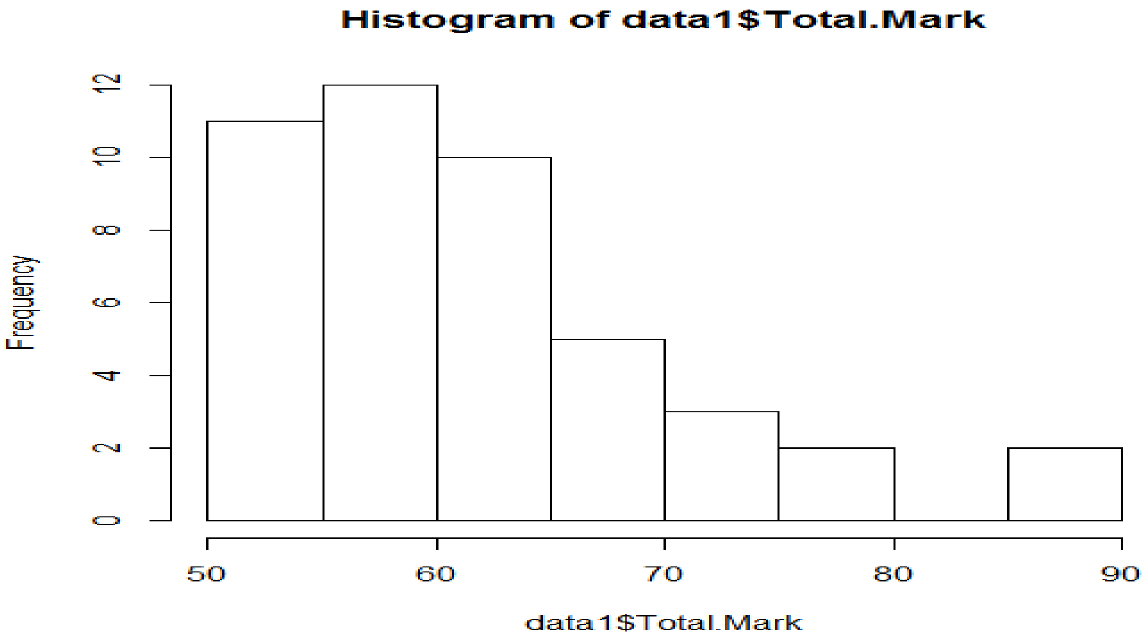


Figure13. *Histogram for Total mark column in given data with R*

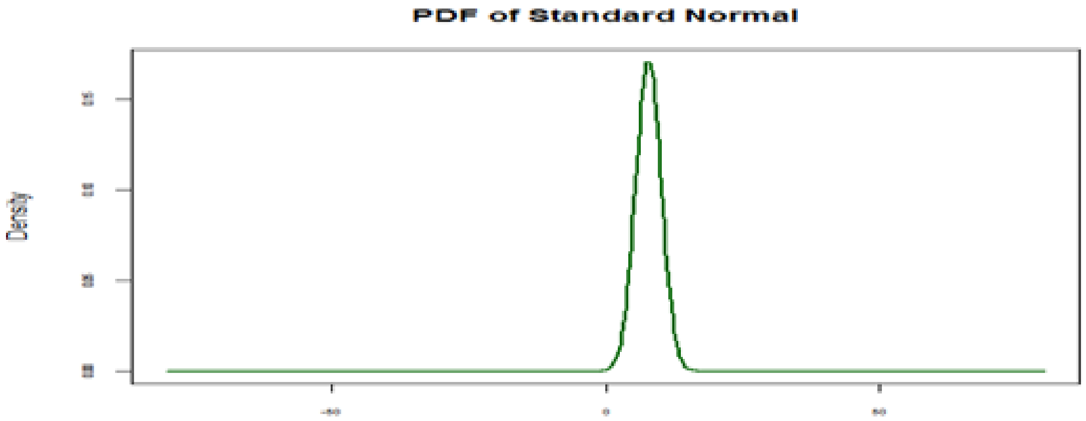


Figure14. *Bell curve on the basis of random numbers with R*

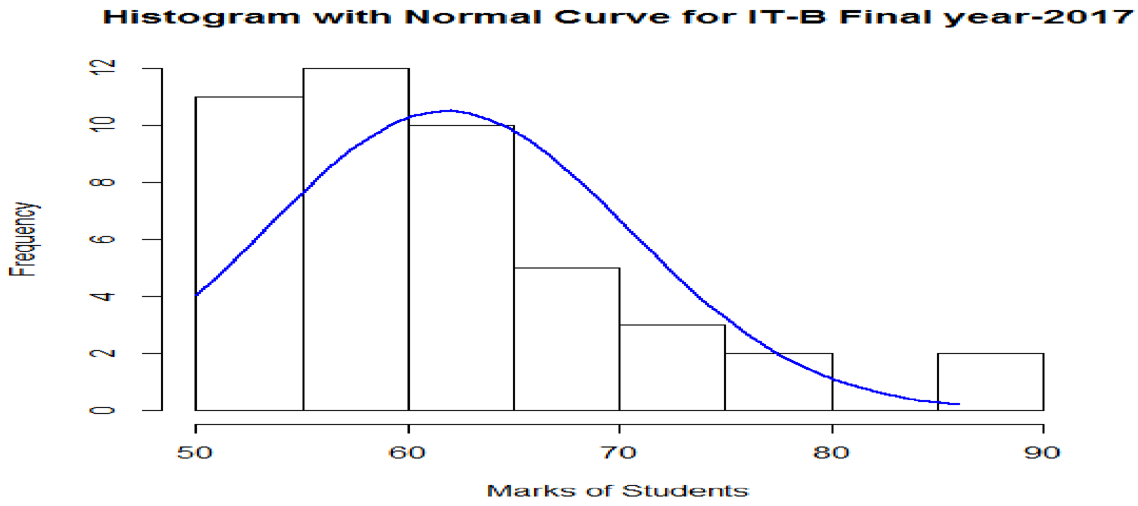


Figure15. *Histogram with Bell curve for the given data(with R)*

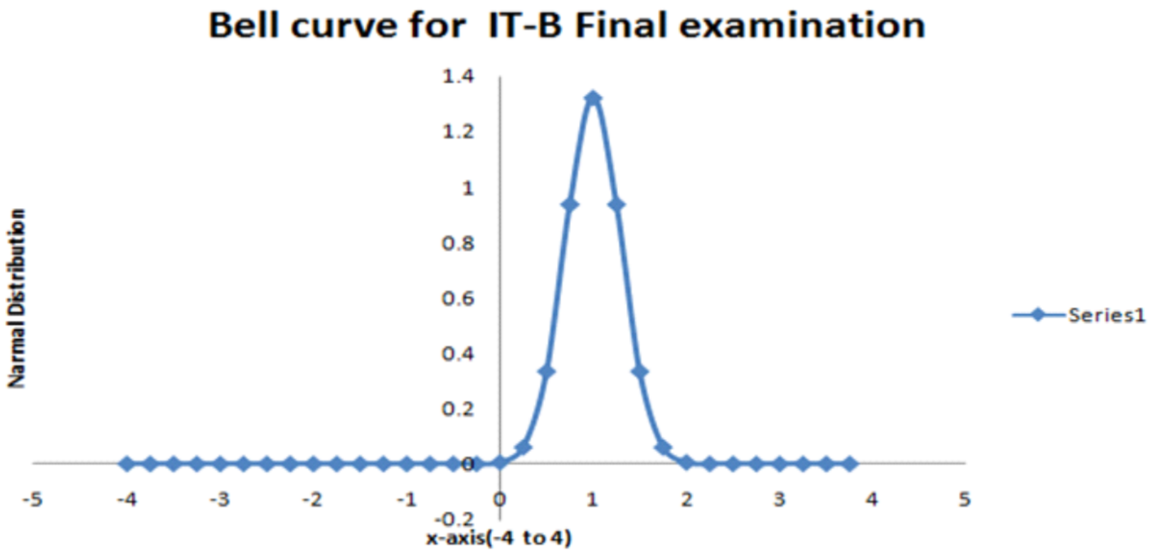
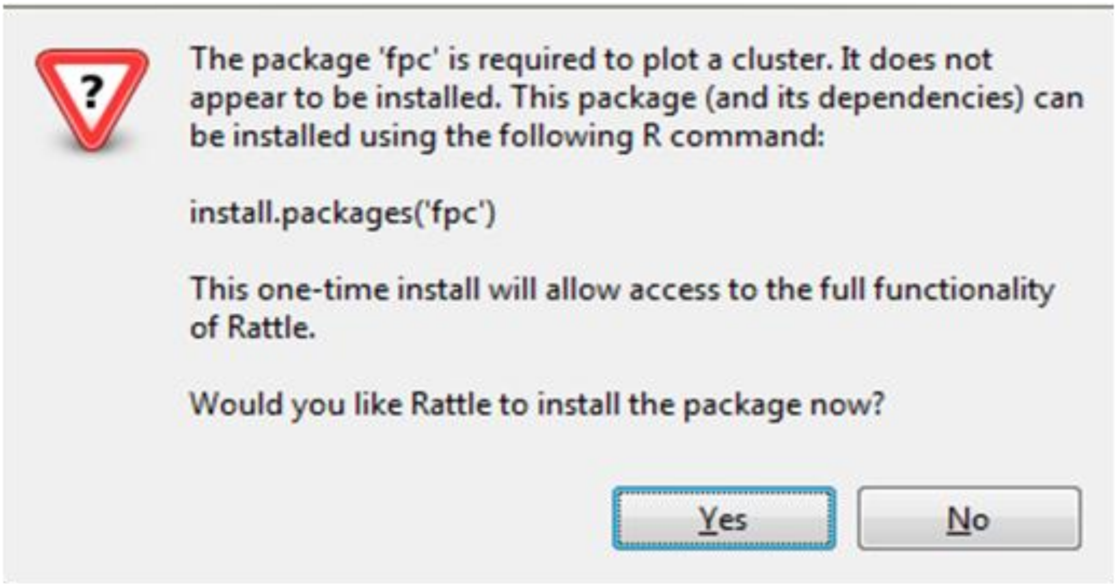


Figure16. *Bell curve for mean=1 and sd-0.3 with MS-Excel 2007*

The “Kurtosis” for given data is examined and noted in the above figures(14,15 & 16). It is noted that from figures 14 & 16 have more peak values than in figure 15. Since figure 14 and 16 are independent in bins where as figure 15 has 10 bins. So as per research community opinion it is not advisable to provide “Hologram for Bell curve” as it is depend upon bins. The authors found that the “Normal distribution” applet is found on [24] for reference and reading purpose. The authors examined the data with that applet and they observed that the “Kertosis” has platikurtic nature and not shown in the figure due to copy right restriction. The readers can refer [21] for more details on different types of shapes of “Kertosis”. From figure-14 & 16 it is observed that both left and right tails are available for “Normal Distribution” of given data set.

The left tail gives the probability of <mean values and the right tail gives the probability of > mean values of given data and it is also called as z-distribution. Generally all are using Fisher-Pearson coefficient of Skewness and not on the basis of Galton Skewness(based on Quartiles)[27].

During the installation of R-Rattle the authors faced some technical issues. Since the R has lot of packages which are necessary for their individual needs. The R software automatically asks the required packages during the execution of required tasks. For example the authors faced the technical problem or requirement for for plot a cluster. The R-Rattle software or tool need “fpc” package for plot the cluster for given dataset as shown in the below figure.



The package 'fpc' is required to plot a cluster. It does not appear to be installed. This package (and its dependencies) can be installed using the following R command:

install.packages('fpc')

This one-time install will allow access to the full functionality of Rattle.

Would you like Rattle to install the package now?

Figure17. *Shows that sample package requirement for cluster plots in R-Rattle*

4. Conclusion

In this paper the authors discussed the different forms of data and its nature at different levels and domains. It is very difficult the exact nature of data with real time example as it is effected with lot of dynamic factors .Generally we are approaching the data understanding on need and context basis. Readers are carefully interpret the meaning of data and its distribution at given interval or its scale values. As a Computer Science student or IT based reader are expected to understand the data and fix the dimensions of the data as per the technical laboratory manuals. Mathematical and Statistical modeling are helpful to analyse and understand the data with different dimensions. But they are supportive only. Suitable Programming Languages and software tools have been properly have been by either software engineers or research community of Computer Science and Engineering Algorithm is necessary to understand exactly to know about the logical flow of data with respect to domain.

Acknowledgements

My special thanks to all my teachers from school level to University level.

References

[1] D Turban, E., Volonino, L., Wood, G. R. & Sipior, J. C. Information technology for management. Hoboken, NJ: J. Wiley & Sons(2013).

[2] Srinatha Karur, Ramana Murthy M.V, International Journal of Sciences: Basic and Applied Research (IJSBAR). Vol- 7, Pp 5-8.

[3] https://en.wikipedia.org/wiki/Decision_theory

[4] V. Kumar, Werner J. Reinartz , Customer relationship management: a database approach, Wiley, Business & Economics,(2006)

[5] David Kroenke, Deborah Bunker, David Wilson, Experiencing MIS, Frenchs Forest publications, N.S.W. Pearson Australia, 2014, ch. 1, pp. :216-217.

[6] http://mathworks.com/products/matlab

[7] https://en.wikipedia.org/wiki/Dimension_(vector_space)

[8] Sanjeev Arora and Boaz Barak, Computational Complexity: A Modern Approach, Princeton University, complexitybook@gmail.com, January 2007,pp:27-28.

[9] Jaiwei Han, Micheline Kamber, Jaian Pei, Data mining concepts and technologies, 3rd edition,pg:227-231.

[10] https://en.wikipedia.org/wiki/Lattice_(order)#/media/File:Hasse_diagram_of_powerset_of_3.svg

[11] http://www.math.hawaii.edu/~ralph/Preprints/latdrawing.pdf

[12] Suhail Owais, Petr Gajdoˇs and V´aclav Sn´aˇsel Department of Computer Science, VSB - Technical University of Ostrava, ˇ tˇr. 17. listopadu 15, 708 33 Ostrava-Poruba Czech Republic Suhail.Owais@vsb.cz Petr.Gajdos@vsb.cz Vaclav.Snasel@vsb.cz

[13] 24th DAAAM International Symposium on Intelligent Manufacturing and Automation, 2013

[14] Formal Concept Analysis – Overview and Applications, Frano Škopljanac-Mačina*, Bruno Blašković, University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb 10000, Croatia

[15] https://www.statmethods.net/advgraphs/trellis.html

[16] Robert I. Kabacoff , R in Action, Second Edition Data analysis and graphics with R, May 2015 , ISBN 9781617291388

[17] https://livebook.manning.com/#!/book/r-in-action-second-edition/chapter-23

[18] https://livebook.manning.com/#!/book/r-in-action-second-edition/chapter-1/149

[19] http://www.sthda.com/english/wiki/lattice-graphs

[20] Mike Biere ,Business Intelligence for the Enterprise, Publisher: IBM Press, Release Date: June 2003, ISBN: 0131413031, pg:20-30.

[21] https://analystprep.com/cfa-level-1-exam/quantitative-methods/kurtosis-and-skewness-types-of-distributions/

[22] https://www.researchgate.net/post/What_is_the_acceptable_range_of_skewness_and_kurtosis_for_normal_distribution_of_data

[23] http://homepage.divms.uiowa.edu/~mbognar/applets/normal.html

[24] https://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html

[25] http://data-mining-tutorials.blogspot.com/

[26] http://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm

AUTHORS' BIOGRAPHY

Dr. Srinatha Karur, has been completed his P.G during the academic year 1994-97 and completed PhD during the year 2012. He has been successfully completed all his tenure periods in terms of academic and technical in charge roles. At present he is at Department of Informatics, Wollega University, Nekemte, Ethiopia as Assistant professor. Along with Data mining his area of interest is Software Engineering , Big data, and Operations Research. He has two decades of academic domain experience including foreign assignments also.

Dr. M. V. Ramana Murthy, is working as Professor of Computer Science, Department of Mathematics and Computer Science, Osmania University, Hyderabad, India. He has 28+ years experience at University level and completed lot of foreign assignments as eminent professor. Computer Networks, Data mining, Big data is area of interest subjects. The professor has tremendous knowledge in Mathematical modeling which is the main key of the Doctorate level research. The professor has tremendous explore on different domains also and completed lot of foreign assignments also.

Citation: *Dr. Srinatha Karur & Prof. M. V. Ramana Murthy (2017). Understanding the Dimensions of Data on Domain Basis, International Journal of Research Studies in Computer Science and Engineering (IJRSCSE), 4(4), pp.53-63, DOI: http://dx.doi.org/10.20431/2349-4859.0404007*