# R: An Emerging Statistical Data Mining Tool

Pooja Kshirsagar
Student ME
Department Of Computer Science and Engineering,
Walchand Institute of Technology, Solapur, India

A.R. Kulkarni
Associate Professor
Department Of Computer Science and Engineering,
Walchand Institute of Technology, Solapur, India

## ABSTRACT

On account of incremental growth in big data analytics, various fields of research and industries require effective data mining tools to derive relevant infsormation from various databases. Thus data mining, big data, machine learning algorithms are all linked with each other and work for a common cause i.e. information. Big Data are very complex in nature and thus mining them is not an easy job. Thus the need of effective data mining tools comes into picture. This paper explores the aspects of R and R studio along with the overview of big data and data mining.

R provides different dimensions to statistical analysis of data sets. However in this paper we discuss the overview of the R studio and demonstrate the implementation of k-means algorithm.(Burda, 2015)

## Keywords

R Tool, R Script, Big data, K-means, Weka, Rapid Miner

## 1. INTRODUCTION

The word “Big data” is buzzing around in almost every sector of today’s era. Various sources of data are all contributing to the enormous diction of data generated by various devices resulting in difficulty in processing and its storage. Thus mining of knowledge and relevant information from this large repository is the major concern. Different structures of different data sizes are handled with the help of various analytical tools and mining techniques.(Pandey et al., 2015)

R is one of these tools which serve an excellent quality of data analysis. This paper studies the R language and R studio introductory sections along with the study of correlation of big data and data mining. This paper is divided into 3 sections. Section II consists of introductory part of big data and Data mining which gives its overview. Section III gives the History and evolution of R language. Section IV gives the brief study of R studio including the installation procedures.(Kosorus et al., 2011)

## 2. AN INTRODUCTION TO THE BIG DATA AND DATA MINING SCENARIO

Big Data was first discovered in the early days of 2012, when unexpected amount of tweets were recognized during the Debate session of the President and Governor. At that instance of time handling of such a huge amount of data was not planned and hence gave rise to the thought of Big Data Analysis. Also various social networks contribute as the Big Data applications since the data or images gathered is huge. Analysing this Big Data, filtering the knowledge and thus deriving meaningful predictions gave rise to various analytical and data mining tools. Novel technologies are being studied across the world and different features of big data are being analysed by researchers.

The sources of big data are a mixture of large-volume, Independent as well as dependent and heterogeneous sources which may be in the centralized or distributed form. In short Big Data is enormously growing data which may have variety of data formats and a high capacity of multiplicity.

### 2.1 Analysis of Big Data

It usually requires computing of higher performance in order to carry out the analysis in limited time. For this, one of the solutions is that the processes are parallelized in order to be executed on the distributed platform, to improvise the time of data analysis. This requirement can be satisfied with the use of Cloud Computing. The cloud is always more efficient than the distributed structures such as grid in many aspects. Hence in the analysis of big data, cloud computing plays a vital role in executing various tasks.

### 2.2 Data Mining Overview

Data mining is the critical step for discovery of knowledge from the data sets. It combines all the analysis procedures that are essential to extract new and relevant information to the end users. Thus for extracting relevant data from the datasets, it is important to prepare the data in the suitable form. Later this prepared data can go through various models to extract results. Another important aspect is that of visualization of the results which is achieved with the help of various statistical data mining tools.

### 2.3 Free Tools for Data Ming

Since past 20 years many software tools have been developed for the data mining. The intension behind these publicly available tools is to provide analysis process for complex data and to provide free platform for data analysis instead of the commercial ones. This is made possible by providing Integrated Development Environments and interesting packages combined with the standard programming languages, which are often freely available.

Among such freeware data analysis tools, which have proved to be very effective in past few years, some of them have managed to stand out even among the commercial ones. Few of them are Rapid Miner, R, Weka, KNIME (all are open source). Almost all of these open source tools have implementations of various general data mining tasks, such as data preparation, data analysis, etc. Also the implementation of various machine learning algorithms in these tools has added to its success rate. Some tools like Rapid Miner are based on graphical integrated environment which consists of components that provide dragging of data that involves pulling of data in, transforming it and then pushing it further. Average user finds it difficult to improvise the under lied code for these components. Whereas tools like R are just simple extensions of the language being used with the help of different packages and GUI extensions?

## 3. HISTORY OF R

For last fifteen years R has been in the development and is the successor of S, which is a statistical language, developed at Bell Labs in 1970s. The source code of R is written in three languages, namely C++, FORTRAN and in R itself. It is an interpreted language and is opted mainly for matrix calculation. R was originally created by RossIhaka and Robert Gentleman at the University Of Auckland, New Zealand. And hence is aptly named after the first letters of its creators and it is currently developed by the R development Core Team.(Ramirez et al., 2015)

### 3.1 N Features of R

The wide areas covered by R include research, statistical applications in science, social studies, economics, medicine and business. The variety of graphical and statistical techniques provided by R include, Linear – Non linear modelling, Time series analysis, Statistical tests, Classification, Regression, Clustering ,etc. R also allows additional functionality to be added into it by defining new functions. R has features similar to functional and object oriented languages.

R is one of the best tools for statisticians and modelling data. The r Programming language is very versatile, sophisticated and has a very good expressive syntax to work around with the data. Some of the factors of its fame are its powerful graphics capability, easy data manipulation techniques and presentation of data in compelling ways.(Kitcharoen et al., 2013)

### 3.2 R Studio

R studio is an IDE for user to run R in more efficient and easy manner. It is open source software. R compiles and runs on various platforms of UNIX, Mac OS and Windows.

## 4. WORKING WITH THE R ENVIRONMENT

Before R studio, R language from the standard website is to be installed.

### 4.1 Steps for Installing R

1. Go for Site “cran.r-project.org”
2. Select the OS you prefer
3. Select the version as per the requirement
4. (Latest R version is 3.2.3, released on 2015-12-10)
5. Follow the installation procedures and provide the destination folder for the file.

### 4.2 Steps for Installing R Studio

1. Go for http://www.rstudio.com/
2. Download the IDE for desktop or choose option as per the requirement.
3. Follow the installation steps and provide the source location for the software.

### 4.3 R IDE

The R console lacks in providing the various visualization options, hence we opt for the IDE R Studio. The IDE is basically divided into 4 windows,

#### *4.3.1 Console*

It is the area where you can type in commands and see output.

#### *4.3.2 Workspace Tab*

It stores any object, value, function you create in your R script. It shows all the active objects currently been used.

#### *4.3.3 History Tab*

It displays the list of commands used so far. It is helpful during testing and running processes. Here either whole list of history can be saved or you can select the commands as per requirement and send them to R script tab, to keep the track of your current work.

#### *4.3.4 File Tab*

Displays all folders and files in your default workspace as if you were on the PC window.

#### *4.3.5 Plot Tab*

It represents all the graphs and statistical figures. To extract the graph or plot, Click on the “Export” tab where you can save the file as image or PDF format. For handy use of the graphs in Word file, in Plot tab, Click Export =>Copy Plot to Clipboard => select metafile =>use it in the word file.

**Package Tab***:* It provides a list of series of packages or add-ons included in the installation of R studio.

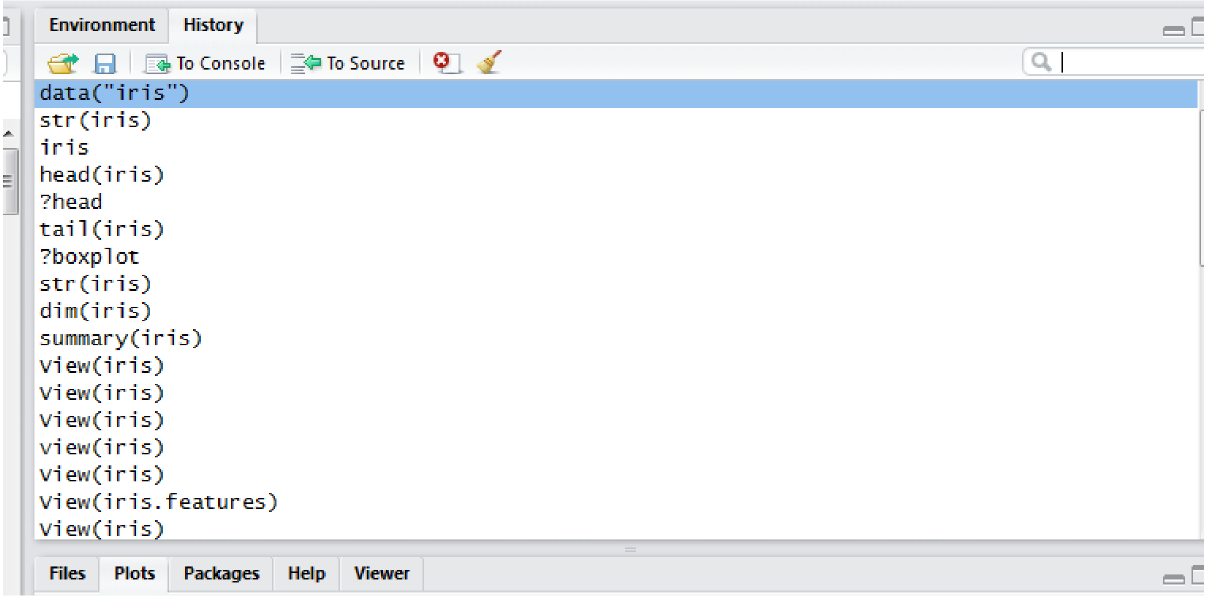
Those Packages (from the list) which are checked is loaded into R, Those which are not, commands related to them won’t work, and you need to select them or type in the console:

#this will check for the foreign package

Library (foreign)

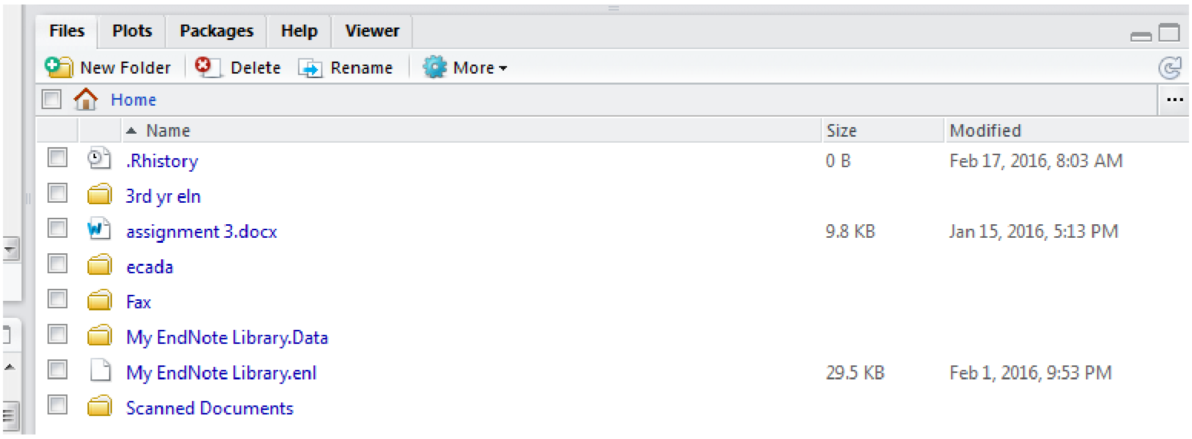
#### *4.3.6 Help Tab*

It gives you brief information about required attribute or function or method



```
data("iris")
str(iris)
iris
head(iris)
?head
tail(iris)
?boxplot
str(iris)
dim(iris)
summary(iris)
View(iris)
View(iris)
View(iris)
view(iris)
View(iris)
View(iris.features)
View(iris)
```

**Fig 1: History Tab**



| | Name | Size | Modified |
|---|---|---|---|
| □ | .Rhistory | 0 B | Feb 17, 2016, 8:03 AM |
| □ | 3rd yr eln | | |
| □ | assignment 3.docx | 9.8 KB | Jan 15, 2016, 5:13 PM |
| □ | ecada | | |
| □ | Fax | | |
| □ | My EndNote Library.Data | | |
| □ | My EndNote Library.enl | 29.5 KB | Feb 1, 2016, 9:53 PM |
| □ | Scanned Documents | | |

**Fig 2: File Tab**

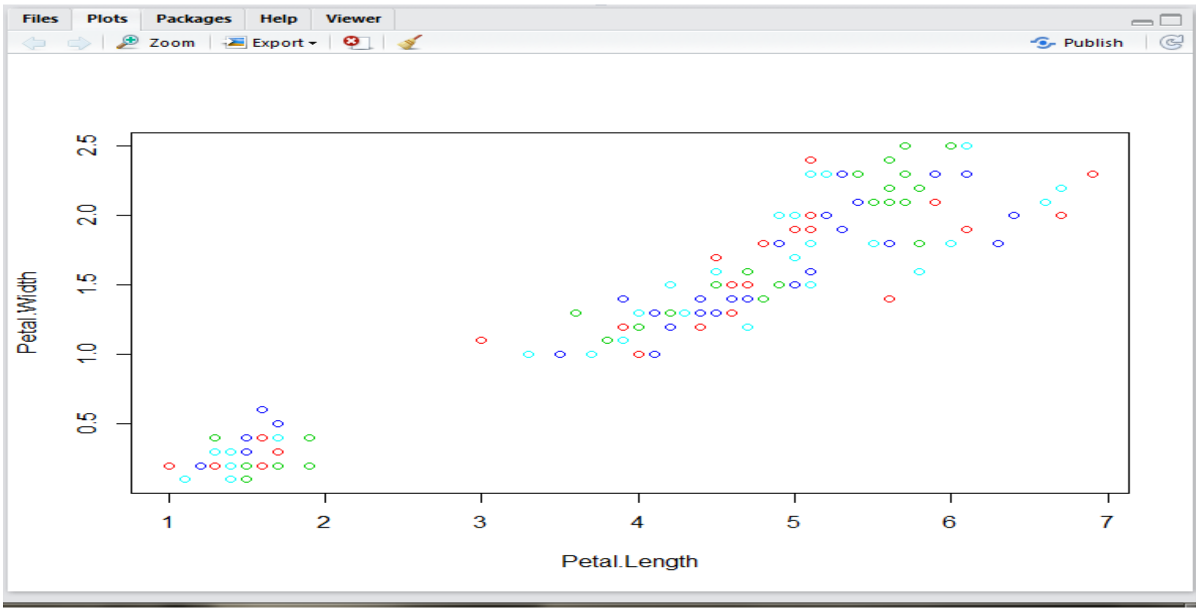
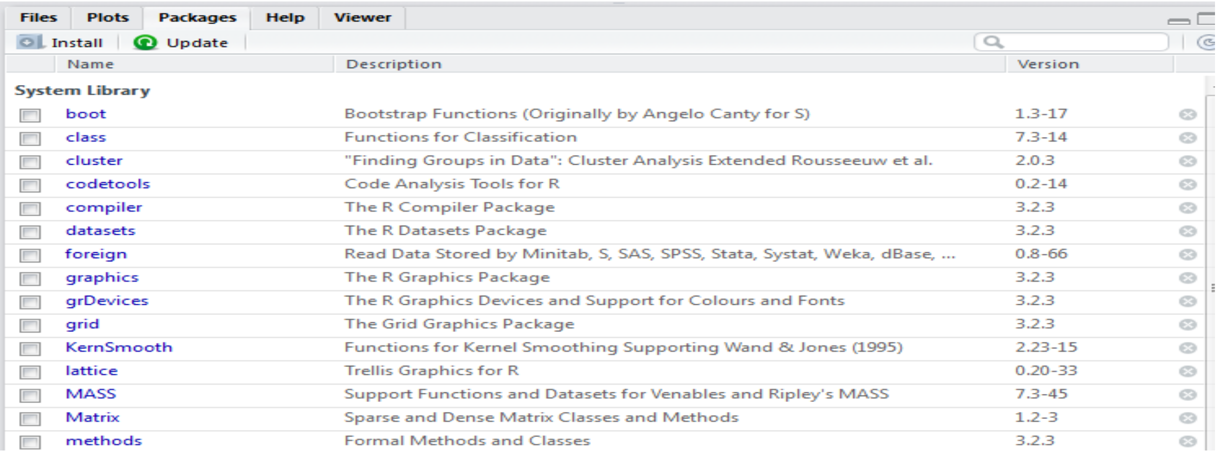
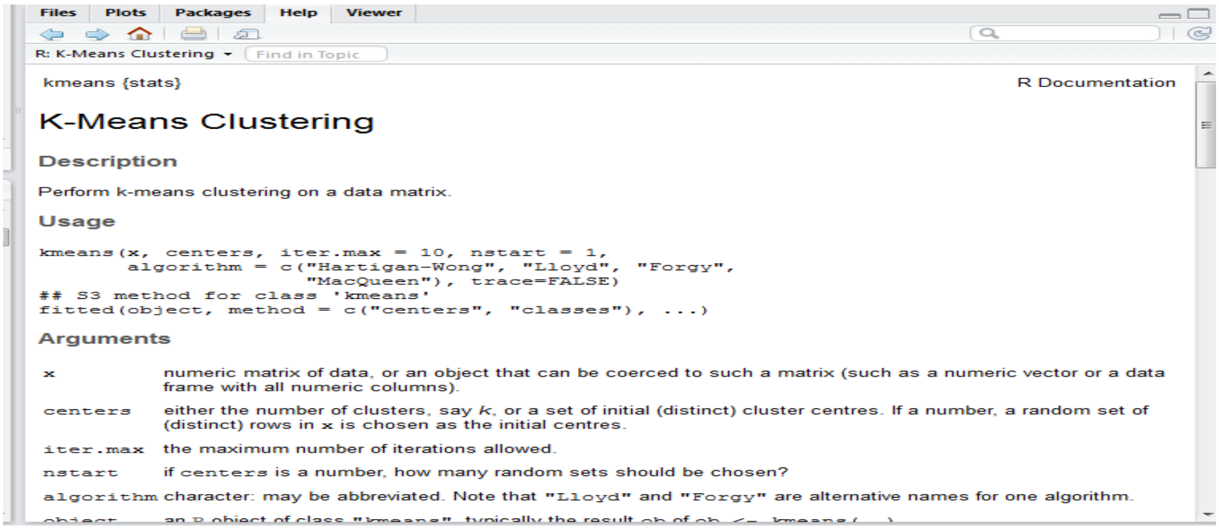


**Fig 3: Plot Tab**



| | Name | Description | Version |
|---|---|---|---|
| | **System Library** | | |
| □ | boot | Bootstrap Functions (Originally by Angelo Canty for S) | 1.3-17 |
| □ | class | Functions for Classification | 7.3-14 |
| □ | cluster | "Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al. | 2.0.3 |
| □ | codetools | Code Analysis Tools for R | 0.2-14 |
| □ | compiler | The R Compiler Package | 3.2.3 |
| □ | datasets | The R Datasets Package | 3.2.3 |
| □ | foreign | Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ... | 0.8-66 |
| □ | graphics | The R Graphics Package | 3.2.3 |
| □ | grDevices | The R Graphics Devices and Support for Colours and Fonts | 3.2.3 |
| □ | grid | The Grid Graphics Package | 3.2.3 |
| □ | KernSmooth | Functions for Kernel Smoothing Supporting Wand & Jones (1995) | 2.23-15 |
| □ | lattice | Trellis Graphics for R | 0.20-33 |
| □ | MASS | Support Functions and Datasets for Venables and Ripley's MASS | 7.3-45 |
| □ | Matrix | Sparse and Dense Matrix Classes and Methods | 1.2-3 |
| □ | methods | Formal Methods and Classes | 3.2.3 |

**Fig 4: Plot Tab**



kmeans {stats} R Documentation

K-Means Clustering

Description

Perform k-means clustering on a data matrix.

Usage

```
kmeans(x, centers, iter.max = 10, nstart = 1,
       algorithm = c("Hartigan-Wong", "Lloyd", "Forgy",
                     "MacQueen"), trace=FALSE)
## S3 method for class 'kmeans'
fitted(object, method = c("centers", "classes"), ...)
```

Arguments

| | |
|---|---|
| x | numeric matrix of data, or an object that can be coerced to such a matrix (such as a numeric vector or a data frame with all numeric columns). |
| centers | either the number of clusters, say *k*, or a set of initial (distinct) cluster centres. If a number, a random set of (distinct) rows in x is chosen as the initial centres. |
| iter.max | the maximum number of iterations allowed. |
| nstart | if centers is a number, how many random sets should be chosen? |
| algorithm | character: may be abbreviated. Note that "Lloyd" and "Forgy" are alternative names for one algorithm. |
| object | an R object of class "kmeans", typically the result ob of ob <- kmeans(..). |

**Fig 5: Plot Tab**

## 4.4 Installing Packages

To install package rgl(which is useful to plot 3D images)

Steps:

1. Click on Install Package icon in the packages tab
2. Write the name of the required package in the pop-up window.
3. Click on install

You will then observe the new package entry in the existing list of the packages.

## 4.5 R script(s) and data view

R script is where you keep a record of your work.

### *4.5.1 To create a new R script following are 3 methods:*

1) File => New => R Script

2) Click on the icon with "+" sign and select "R Script"

3) Ctrl+Shift+N

OR

Type in console:

```
#displays current directory
getwd ()
#sets new directory
setwd ("C: /myfolder/data")
```

## 4.6 Changing the Working Directory:

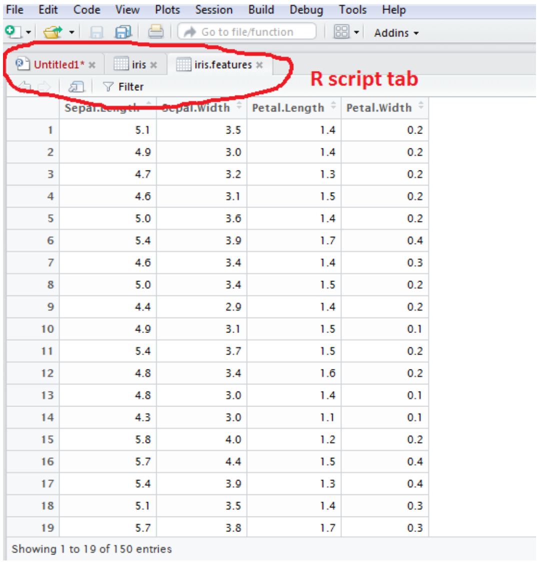
If you have multiple projects, you can change the working directory for your convenience as follows:

From Menu Bar click on Session =>Set Working Directory => Choose Directory => select the folder.

You will then observe the new package entry in the existing list of the packages.

This is useful when your documentation is huge and your working directory has many files. Now handling these files in an efficient manner is the real trick.

Also if your different working directory have some common part then you can immediately import any required data among the inter – directory level by managing them in an efficient manner.

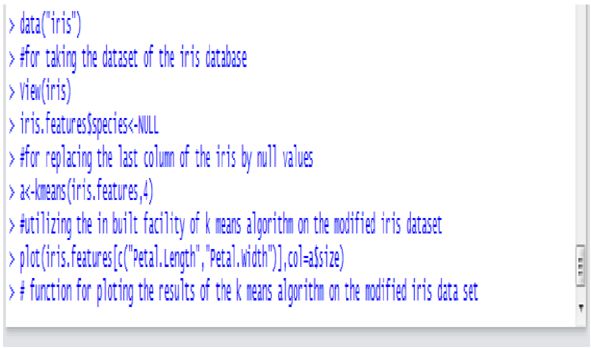


| | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width |
|---|---|---|---|---|
| 1 | 5.1 | 3.5 | 1.4 | 0.2 |
| 2 | 4.9 | 3.0 | 1.4 | 0.2 |
| 3 | 4.7 | 3.2 | 1.3 | 0.2 |
| 4 | 4.6 | 3.1 | 1.5 | 0.2 |
| 5 | 5.0 | 3.6 | 1.4 | 0.2 |
| 6 | 5.4 | 3.9 | 1.7 | 0.4 |
| 7 | 4.6 | 3.4 | 1.4 | 0.3 |
| 8 | 5.0 | 3.4 | 1.5 | 0.2 |
| 9 | 4.4 | 2.9 | 1.4 | 0.2 |
| 10 | 4.9 | 3.1 | 1.5 | 0.1 |
| 11 | 5.4 | 3.7 | 1.5 | 0.2 |
| 12 | 4.8 | 3.4 | 1.6 | 0.2 |
| 13 | 4.8 | 3.0 | 1.4 | 0.1 |
| 14 | 4.3 | 3.0 | 1.1 | 0.1 |
| 15 | 5.8 | 4.0 | 1.2 | 0.2 |
| 16 | 5.7 | 4.4 | 1.5 | 0.4 |
| 17 | 5.4 | 3.9 | 1.3 | 0.4 |
| 18 | 5.1 | 3.5 | 1.4 | 0.3 |
| 19 | 5.7 | 3.8 | 1.7 | 0.3 |

Showing 1 to 19 of 150 entries

**Fig 6: Script Tab**

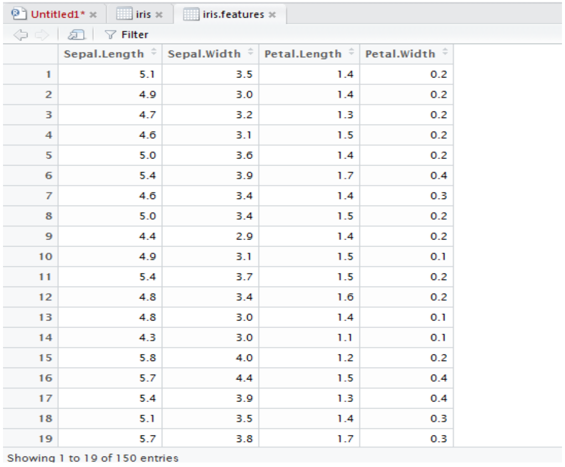
## 5. IMPLEMENTATION OF K MEANS ALGORITHM WITH THE HELP OF TRAINING DATASET OF IRIS



```
> data("iris")
> #for taking the dataset of the iris database
> View(iris)
> iris.features$species<-NULL
> #for replacing the last column of the iris by null values
> a<-kmeans(iris.features,4)
> #utilizing the in built facility of k means algorithm on the modified iris dataset
> plot(iris.features[c("Petal.Length","Petal.Width")],col=a$size)
> # function for ploting the results of the k means algorithm on the modified iris data set
```

**Fig 7: Code for K-Means Algorithm**

The data set of IRIS is being trained with some nominal functions and then plot function is implemented on



| | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width |
|---|---|---|---|---|
| 1 | 5.1 | 3.5 | 1.4 | 0.2 |
| 2 | 4.9 | 3.0 | 1.4 | 0.2 |
| 3 | 4.7 | 3.2 | 1.3 | 0.2 |
| 4 | 4.6 | 3.1 | 1.5 | 0.2 |
| 5 | 5.0 | 3.6 | 1.4 | 0.2 |
| 6 | 5.4 | 3.9 | 1.7 | 0.4 |
| 7 | 4.6 | 3.4 | 1.4 | 0.3 |
| 8 | 5.0 | 3.4 | 1.5 | 0.2 |
| 9 | 4.4 | 2.9 | 1.4 | 0.2 |
| 10 | 4.9 | 3.1 | 1.5 | 0.1 |
| 11 | 5.4 | 3.7 | 1.5 | 0.2 |
| 12 | 4.8 | 3.4 | 1.6 | 0.2 |
| 13 | 4.8 | 3.0 | 1.4 | 0.1 |
| 14 | 4.3 | 3.0 | 1.1 | 0.1 |
| 15 | 5.8 | 4.0 | 1.2 | 0.2 |
| 16 | 5.7 | 4.4 | 1.5 | 0.4 |
| 17 | 5.4 | 3.9 | 1.3 | 0.4 |
| 18 | 5.1 | 3.5 | 1.4 | 0.3 |
| 19 | 5.7 | 3.8 | 1.7 | 0.3 |

Showing 1 to 19 of 150 entries

**Fig 8. Trained IRIS Datase**

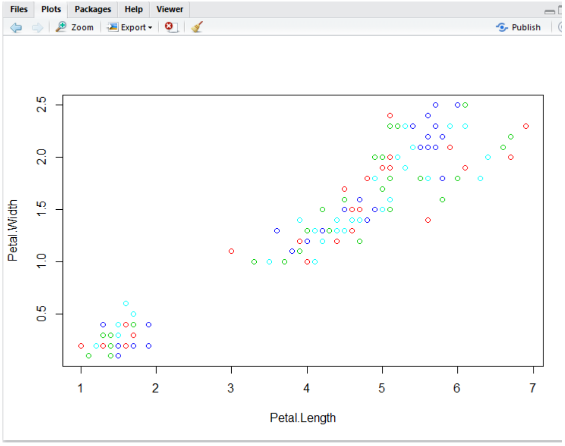


**Fig .9 Plot of the Trained Data**

## 6. CONCLUSION

Big Data is a Combination of information which is in heterogeneous, time varying, uncertain and redundant form. Tools that will help improvise this analysis are of great importance. This paper explores the different aspects involved in Big Data and an overview on the data mining. Also the statistical tool for data analysis: R is described along with the implementation of the k- means algorithm in the R IDE. Since R is an emerging Statistical Tool many upcoming developments in it and R shiny package will allow our data to get deployed over the web in more effective manner.

## 7. ACKNOWLEDGMENTS

My thanks to my guide Prof. A. R. Kulkarni who has contributed towards development of this research.

## 8. REFERENCES

[1] BURDA, M. Linguistic fuzzy logic in R. Fuzzy Systems (FUZZ-IEEE), 2015 IEEE International Conference on, 2-5 Aug. 2015. 1-7.

[2] KITCHAROEN, N., KAMOLSANTISUK, S., ANGSOMBOON, R. & ACHALAKUL, T. RapidMiner framework for manufacturing data analysis on the cloud. Computer Science and Software Engineering (JCSSE), 2013 10th International Joint Conference on, 29-31 May 2013 2013. 149-154.

[3] KOSORUS, H., HONIGL, J. & KUNG, J. Using R, WEKA and RapidMiner in Time Series Analysis of Sensor Data for Structural Health Monitoring. Database and Expert Systems Applications (DEXA), 2011 22nd International Workshop on, Aug. 29 2011-Sept. 2 2011 2011. 306-310.

[4] PANDEY, R., SRIVASTAVA, N. & FATIMA, S. Extending R Boxplot Analysis to Big Data in Education. Communication Systems and Network Technologies (CSNT), 2015 Fifth International Conference on, 4-6 April 2015 2015. 1030-1033.

[5] RAMIREZ, C., NAGAPPAN, M. & MIRAKHORLI, M. Studying the impact of evolution in R libraries on software engineering research. Software Analytics (SWAN), 2015 IEEE 1st International Workshop on, 2-2 March 2015 2015. 29-30.

[6] http://dss.princeton.edu/training/RStudio101.pdf

[7] https://www.youtube.com/watch?v=sAtnX3UJyN0