International Journal of Trendy Research in Engineering and Technology
IJTRET
Volume 4 Issue 1 Feb'2020
ISSN NO 2582-0958

# Implementation of Data Mining Concepts in R Programming

J. Umarani[1]* and S. Manikandan[2]
[1] Research Scholar, Research and Development Centre, Bharathiyar University, Coimbatore,
Email:umashenthaan@gmail.com
[2]Professor and Head, Department of Computer Science and Engineering,
Sri Ram Engineering College, Chennai – 602024, Tamil Nadu, India.
Corresponding Author :manidindigul@rediffmail.com

**Abstract**: - Data Mining is a process used to extract the usable data from a larger set of any raw data. It involves analyzing data patterns in large batches of data using one or more software. R is a programming language for the purpose of statistical computations and data analysis. The R language is widely used by the data miners and statisticians on high dimensional pattern extraction. R is freely available under the GNU General Public Licenses and the source code is written in FORTAN, C and R.It is a GNU project. The pre-compiled binary versions are freely available for various flavours of operating system. R is basically command line interface (CLI) and various GUI interfaces are also available now a day. In this articlefocuses the concept of R like; getting data into and out of R and packages related to data mining and data visualization.

**Index Terms:**
***Data mining, R Programming, Packages for Data mining, Data sets,Data Visualization.***

## I.INTRODUCTION

### A. Introduction to Data mining

Data mining is a set of techniques and methods relating to the extraction of knowledge from large data sets [through automatic or semi-automatic methods] and further scientific, industrial or operational use of that knowledge. DM is closely related to the statistics as an applied mathematical discipline with an analysis of data that could be defined as the extraction of useful information from data. As an application of data mining, business can learn more about their customers and develop more effective strategies related to various business functions and in turn leverage resources in a more optimal and insightful manner. It helps business be closer to their objective and make better decisions.

Data mining involves effective data collection and warehousing as well as computer processing. For segmenting the data and evaluating the probability of future events, data mining uses sophisticated mathematical algorithms. Data mining is also known as Knowledge Discovery in Databases (KDD).

***Features of Data mining:***

- ✓ Automatic pattern prediction based on trend and behavior analysis.
- ✓ Prediction based on likely outcomes.
- ✓ Creation of decision-oriented information.
- ✓ Focus on large data sets and databases for analysis.
- ✓ Clustering based on finding and visually documented groups of facts not previously known.
- ✓ The main techniques for data mining include classification and prediction, clustering, outlier detection, association rules, sequence analysis and text mining, social network analysis and text mining, and also some new techniques such as social network analysis and sentiment analysis [01].

### B. Introduction to R Programming

R is a programming language and an environment for statistical computing and it is similar to the S language developed at AT&T Bell Laboratories by Rick Becker, John Chambers and Allan Wilks. There are versions of R for the Unix Windows and Mac families of operating systems. Moreover, R runs on different computer architecture like Intel, Alpha systems, PowerPC and Sparc system.

R was initially developed by Ihaka and Gentleman (1996) both from the University of Auckland, New Zealand. The current development of R is carried out by a core team of a dozen people from different institutions around the world. R development takes advantage of a growing community that cooperates in its development due to its open source philosophy. In effect, the source code of every R component is freely available for inspection and /or adaption. This fact allows you to check and test the reliability of anything you use in R.

Data analysis with R is done in a series of steps; programming, transforming, discovering, modelling and communicate the results.

**C.R Studio**

RStudio is an integrated development environment (IDE) for R and can run on various operating systems like windows, Mac OS X and Linux. It is a very useful and powerful tool for R programming. When RStudio is launched for the first time, you can see a window similar to figure 1. There are for panels:

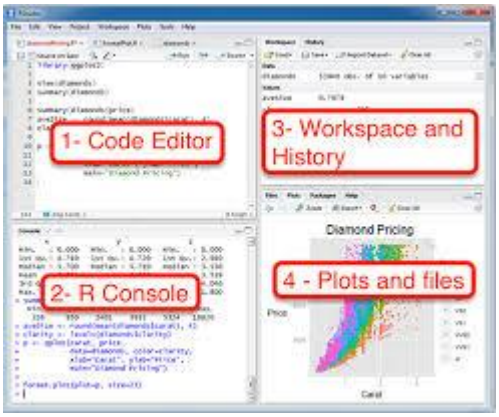


Figure.1. RSudio Desktop Window

**1**. Source panel (top left), which shows your R source code. If you cannot see the source panel, you can find it by clicking menu "File", "New File", and then "R Script". You can run a line or selection of code by clicking the "Run" bottom on top of source panel, or pressing "Ctrl+Enter".

**2.** Console Panel (bottom left), which shows outputs and system messages displayed in a normal R Console;

**3.** Environments/History/Presentation panel (top right), whose three tabs shows respectively all objects and function loaded in R, a history of submitted R code, and Presentations generated with R.

**4.**Files/Plots/Packages/Help/Viewer panel (bottom right), whole tabs show respectively a list of files, plots, R packages installed, help documentation and local web content

**How to create new project in R Studio?**

Step 1: Click Project button at the top right corner.
Step 2: Choose New Project
Step 3: Select create project from new directory and then "Empty Project". Then type the name of the directory and click create project button.

**Folder in RStudio Environment**

There are three main folders are created automatically after creating the new project from Rstudio; there are listed as below;

1. Code- where to put your R code
2. Where to put your datasets; and
3. Figures-where to put produced diagrams

**Additional Folder in RStudio Environment**

1. Rawdata-where to put all raw data
2. Models where to put all produced analytics models, and
3.Reports- where to put your analysis reports.

## II.RELATED WORKS

Concepts of data mining with R programming are discussed by many researchers here some of them are considered for this research article.

YanchangZhao [01] published a book R and Data Mining: Examples and Case Studies, which consists fifteen chapters they are; introduction, Data Import and Export, Data exploration, Decisions Trees and Random Forest, Regression, Clustering, Outlier Detection, Time series analysis mining, Association Rules and Text Mining, Social Network Analysis, remaining are case studies and online resources. Each chapter delivers the data mining concepts and related packages and methods are discussed with example. From this book learned more R programing concepts related to Data mining.

Vipin Kumar [02] published a book, Data Mining with R Learning with case studies, Data Mining and Knowledge Discovery Series, which consists five chapters named introduction, Predicting Algae Blooms, Predicting Stock Market Returns, Detecting Fraudulent Transactions and Classifying Microarray Samples. Each chapter delivers the data mining concepts with their case study applications. Research problem identified from this reference.

Miss. TejashreeU.Sawant [03] presented an article R: Data Mining Tool and Its Applications, which delivers the concepts of the Data mining tool and applications in R. Six open source data mining tools and its descriptions are listed in tabulated form. The tool names are listed as follows; RapidMiner, WEKA, R, Orange, KNIME, NLTK. R programming concepts are implemented in various applications; some of the applications are listed below;

a) Chemometrics and Computational Physics
b) Clinical Trial Design, Monitoring, and Analysis
c) Computational Econometrics
d) Analysis of Ecological and Environmental Data

e) Design of Experiments (DoE) & Analysis of Experimental Data
f) Empirical Finance
g) Statistical Genetics
h) Medical Image Analysis
i) Natural Language Processing (NLP)
j) Official Statistics & Survey Methodology
k) Analysis of Pharmacokinetic Data
l) Phylogenetic, Especially Comparative Methods
m) Psychometric Models and Methods
n) Reproducible Research
o) Statistics for the Social Sciences

Sonja Pravilovice [04] presented an article R Language in data mining techniques and statistics, the main aim of this study was to point to the application of modern programming langue's and statistical packages without which modern science and research work in many areas of economics, finance, medicine, meteorology, engineering, and data mining cannot be imagined today. Application of R as a programming language and statistical software is much more than a supplement of Stata, SAS, and SPSS. Although it is more difficult to learn, the biggest advantage of R is its free-of-charge feature and the wealth of specialized application packages and libraries for a huge number of statistical, mathematical and other methods.

R is a simple, but very powerful data mining and statistical data processing tool and once "discovered", it provides users with an entirely new, rich and powerful tool applicable in almost every field of research.

Sadiq Hussain [05] presented article for educational data mining using R Programming and R Studio, the main objective of this article is to analyses the performance of B.A. students of Dibrugarh University with respect to the caste and gender. The analysis of variance is a commonly used to determine difference between several samples. R provides a function to conduct ANOVA so: aov (model, data). The first stage is to arrange your data in a .CSV file. Use a column for each variable and give it a meaningful name. Second stage is to read your data file into memory and give it a sensible name. The next stage is to attach the data set to that the individual variables are read into memory. Finally it is required to define the model and run the analysis.

From statistical analysis, it is confirmed that the OBC students' performance are better than the other category students. The results of the candidates further analyzed by grouping as the First Class Candidates and Second Class Candidates. It was found that the statistical difference between male and female candidates and among the caste of the candidates were significant because of the results of the Second Class students.

## III.METHODOLOGY

Proposed methodology is organized in the sequence of steps are depicted in the figure.1. Architecture of proposed methodology.

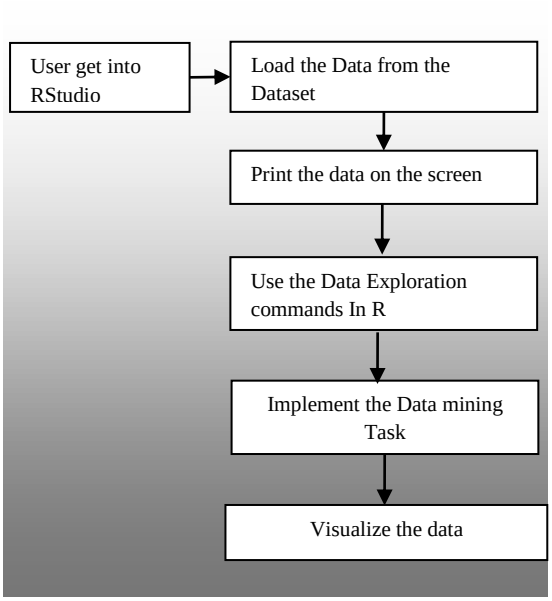


Figure 1: Architecture of the proposed methodology

### A. Import Dataset

Weather dataset is downloaded from the internet and loaded to the Studio in Two aspects; the first one is via RStudio menu options; the sequence of steps is as follows;

Click file in RStudio->Import dataset-from text->Select file to import-> choose the file from the location -> click import, and then the available records are displayed in the screen.

In command line mode, we can use the following R code to import the dataset into **RConsole**.

*Weather<-read.csv("E:/weather.csv"*
*Print(Weather)* after this command is executed the all records in the datasets are displayed on the screen.

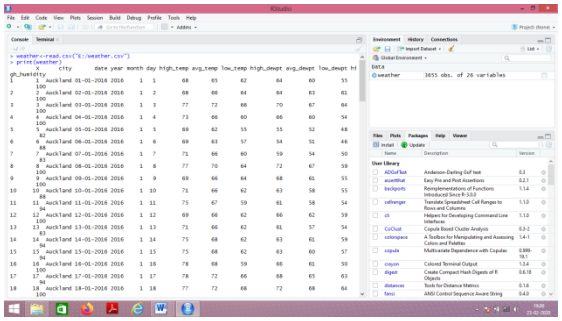

Figure.2: Code and Result Window for read and print command in R Programming.

The above mentioned read command code are executed in R programming and print the results, its depicted in figure2.

## B. Data Exploration Methods in R

The The following table.1.represents the few methods in data exploration in R programming.

Table.1. Data Exploration Methods in R

| *S.No* | *Methods* | *Meaning* |
|---|---|---|
| 01 | Dim() | Display the No.of.Rows and columns |
| 02 | Names()/attributes() | Display the attribute names |
| 03 | Class() | Display the class name for the dataset |
| 04 | Row.names() | Display the row name in the dataset |
| 05 | Head() | Display the top few records in the data set |
| 06 | Summary() | Display the summary of individual attributes in the data set |
| 07 | Var() | Display the values of variable in dataset |

These methods are applied to the weather dataset and run in RStudio the resulting window and code windows are depicted as follows;

*01.Dim(Weather), run in RStudio and print the output as [1] 3655 26*

02. names(weather)/*attributes(weather), run in RStudio and print the output printed, it is in the figure.3.*

*03. class(weather), run in RStudio and print the output printed, [1] "data.frame", it is in the figure.3.*

*04. row.names(weather), run in RStudio and print the output printed, [1] "data.frame", it is in the figure.3*

*05. head (weather), run in RStudio and print the output, [1] "data.frame", it is in the figure.4.*

*06. summary(weather), run in RStudio and print the output, [1] "data.frame", it is in the figure.4.*

*07.*`>var(weather$avg_wind)`

`[1] 14.93843,` *it is in the figure.5.*

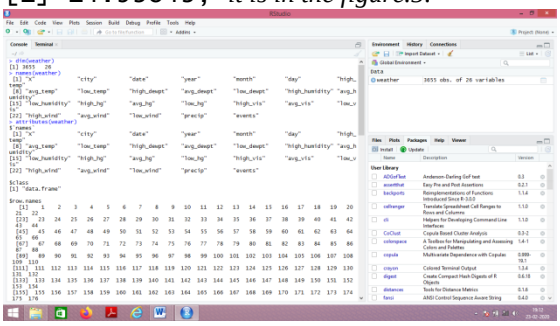

Figure.3. name, attribute and class output window

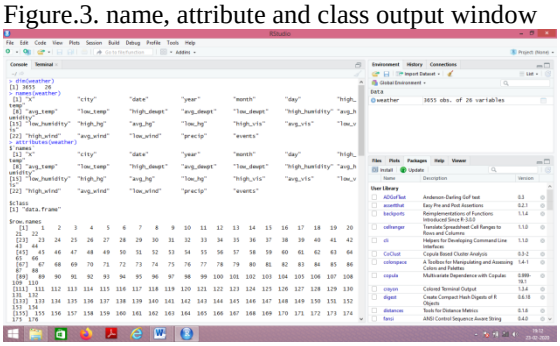

Figure.4. head and summary output window

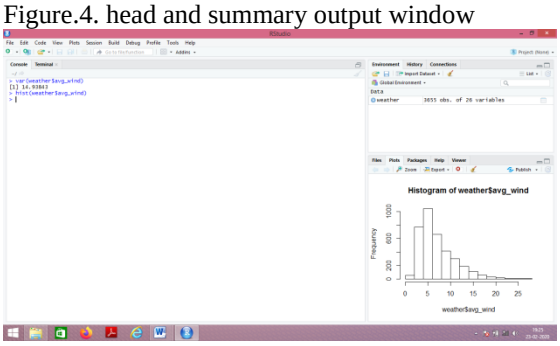


Figure.5. var and hist output window

Mentioned above figure numbers 3,4 and 5 are represented the code and output windows for data visualization methods.

## C.Data Mining Task

K-means clustering algorithm is applied to the mdvis dataset and produces the result.

```
> (kmeans.result<- kmeans(mdvis, 5))
K-means clustering with 5 clusters of sizes 445, 446, 446, 445, 445


Cluster means:
```

```
X numvisit    reform       badh
age     educ     educ1     educ2
educ3   agegrp
1 1115.0 1.7752809 0.5595506
0.13707865 36.65618 2.191011
0.1977528 0.4134831 0.3887640
1.516854
2  223.5 7.4304933 0.5448430
0.10762332 36.54484 2.094170
0.2600897 0.3856502 0.3542601
1.522422
3  669.5 3.0403587 0.5044843
0.09641256 35.53139 2.114350
0.2578475 0.3699552 0.3721973
1.475336
4 1560.0 0.6876404 0.5573034
0.15280899 36.95506 2.060674
0.2831461 0.3730337 0.3438202
1.570787
5 2005.0 0.0000000 0.3640449
0.07415730 38.22022 1.995506
0.2337079 0.5370787 0.2292135
1.687640
       age1      age2      age3
loginc
1 0.6157303 0.2516854 0.1325843
7.719951
2 0.6233184 0.2309417 0.1457399
7.698639
3 0.6681614 0.1883408 0.1434978
7.690957
4 0.5887640 0.2516854 0.1595506
7.679370
5 0.5393258 0.2337079 0.2269663
7.775296

Clustering vector:
   [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
  [50] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
  [99] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [148] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [197] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [246] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [295] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [344] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [393] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [442] 2 2 2 2 2 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [491] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [540] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [589] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [638] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [687] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [736] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [785] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [834] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [883] 3 3 3 3 3 3 3 3 3 3 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [932] 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [981] 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1
[ reachedgetOption("max.print") --
omitted 1227 entries ]
```

```
Within cluster sum of squares by cluster:
[1] 7391207 7462665 7450511 7397835 7401236
 (between_SS / total_SS=  96.0 %)

Available components:

[1] "cluster"      "centers"      "totss"
  "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
```

## IV. RESULTS AND DISCUSSION

Data visualization is one of the best part in the presentation of research documentation; here data exploration methods are applied to the *weather* data set and visualize the results based on attribute performance.

```
>plot(weather$high_temp,weather$date)
```

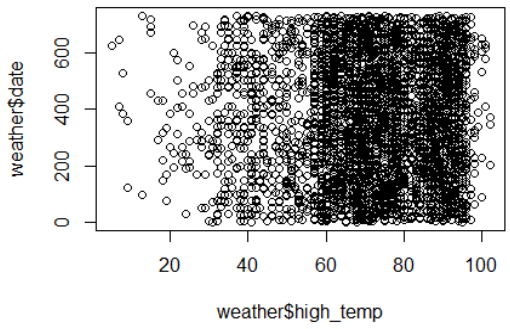


Figure.6: Datewise high temperature

```
>plot(weather$high_humidity,weather$month)
```

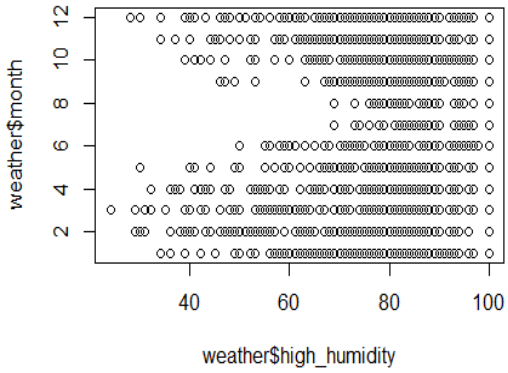


Figure.7: Month wise Humidity

```
>plot(weather$high_hg,weather$month)
```

Mentioned above figure numbers 6 and 7 are represented Date wise high temperature and Month wise Humidity.

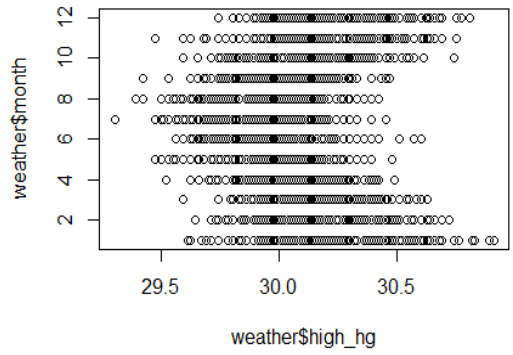


Figure.8: Month wise Hg

```
>plot(weather$high_wind,weather$month)
```

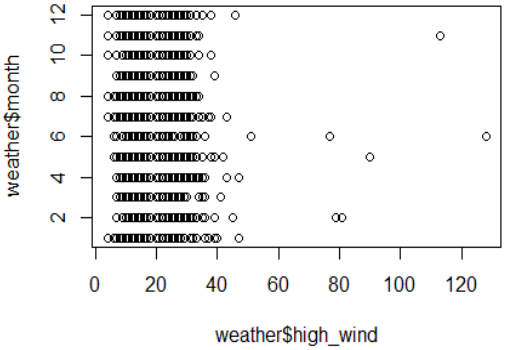


Figure.9: Month wise high wind

Mentioned above figure numbers 8 and 9are represented:Month wise Hgand Month wise high wind.

## V.CONCLUSION

This research article focused on how the data mining concepts are implemented in R programming environment. First section of this article provides the basic concepts of data mining and the R Programming like import of data set and data exploration methods in R programming. Second section discussed on the implementation of data mining task, here the kmeans clustering in R programming are implemented to the mdvis data sets for five cluster grouping. Finally the data visualization are discussed with the weather data set with the help of plot command in R Programming.

## REFERENCES

1. Yanchang Zhao, R and Data Mining: Examples and Case Studies, yanchang@rdatamining.com, http://www.RDataMining.com, April 26, 2013.
2. Vipin Kumar, Data Mining with R Learning with case studies, Data Mining and Knowledge Discovery Series, Chapman & Hall/CRC. Visit the Taylor & Francis Web site at http://www.taylorandfrancis.com
3. Miss. TejashreeU.Sawant, R: Data Mining Tool and Its Applications, International Journal of Advanced Computer Technology & Management (IJACTM), ISSN: 2343-662X, Volume: I, Issue: I May 2016.
4. Sonja Pravilovic, R Language in data mining techniques and statistics, American Journal of Software Engineering and Applications 2013;2(1);7-12. Published online February 20.2013 (http://www.sciencepublishinggroup.com/j/ajsea)doi:10.11648/j.ajsea.20130201.12.
5. Sadiq Hussain, "Educational Data Mining Using R Programming and R Studio", Journal of Applied and Fundamental Sciences, Vol 1(1) April 2015, Conference Paper.