R: An Open Source Software Environment for Statistical Analysis

Yatrik Patel[1], Divyakant Vaghela[2], Hitesh Solanki[3] and Mitisha Vaidya[4]
[1] Scientist – D (CS),[2,3] Scientist – B (CS), [4] Project Officer (CS)

The research data and its efficient analysis play a vital role in successful research as well as in drawing meaningful inferences. Students, scholars, research institutes, Government and non-Government organizations around the World depend heavily on statistical and data analysis to carry out their research as well for taking decisions based on statistical analysis of data. Statistics plays a vital role in almost all domains of knowledge. This article provides an overview of various comprehensive, proprietary and non-proprietary statistical packages that are available in the market place for carrying out statistical analyses on data such as SPSS, Stata, Minitab, SAS, etc. The article especially focuses on non-proprietary and open source statistical package called R, which is widely used in all research domains for statistical analysis, data mining and data visualization across the world.

Introduction

The statistics is extensively used across all major domains of knowledge including engineering & technology, medical & health sciences, chemical sciences, agricultural sciences, biological sciences, physical sciences, social sciences and arts and humanities, etc. While applying research methodology in any knowledge domain, researchers should have basic knowledge of statistical methods along with knowledge on use of statistical packages to carry out meaningful analyses on their research data and draw meaningful inferences. Some of the commonly used statistical methods and their possible order of their application are as follows:

- Measures of central tendencies such as mean, median, mode;
- Measures of dispersions such as range, variance, standard deviations, co-efficient of variance;
- Test of hypothesis for parametric test such as one sample t-test, two independent sample t-test, paired t-test;
- Analysis of variance: one way ANOVA and two way ANOVA;
- Test of hypothesis for non-parametric test such as Kruskal Wallis H test, Mann–Whitney U test, Wilcoxon signed ranks test, Pearson's chi squared test;
- Spearman's correlation coefficient and Pearson's correlation coefficient;
- Simple and multiple regression analysis;
- Logistic regression analysis;
- Time series analysis; and
- Factor analysis

The selection of appropriate statistical model or method depends on the objective of research and nature of data (i.e. qualitative and quantitative). In the fast-growing technological world of Data Science, a number of proprietary and non-proprietary statistical packages like PSPP, Stata, SPSS, SYSTAT, SAS, Minitab, Statistica, SOFA statistics and R, etc. are available in the marketplace that can be deployed for interpreting the results derived from any size and type of data and for drawing statistical inferences. However, some of the statistical packages involve huge cost on its procurement or licensing, whereas some are freely available as open source. R is very popular and widely used with extensive features. It provides an open research environment for statistical analysis. Forthcoming sections of this article briefly explain various statistical packages with emphasis on R statistical package, an open source statistical package used extensively by research communities in various domains of knowledge.

1. Popular Statistical Packages

The statistical and data analytical packages predominantly provide functionalities for organization, manipulation, analyses and graphical representation of research data of all kinds and types irrespective of its knowledge domains. All statistical packages support various data types such as numeric, date and string. Data can be imported from spreadsheets, ASCII text files, csv, tsv, relational database management system and other statistical packages. The statistical packages are used to derive results through statistical methods such as descriptive statistics, bivariate statistics, and prediction for numerical outcomes, identifying groups, simulation modeling, regression analysis, outlier tests, tolerance intervals, stability studies, equivalence tests, design of experiments (DOE), measurement systems analysis, capability analysis, hypothesis tests, and analysis of variance (ANOVA), etc. All statistical packages provides features for representing resultant data as charts and graphs which can be exported as an image in any graphical file format such as jpg, png, gif, etc. The data in tabular format can be exported in various formats including word, excel, tab delimited text, pdf, xml, html, jpg, etc. Statisticians, data scientists and researchers are using the statistical packages as per their requirement.

The most commonly used proprietary and non-proprietary statistical packages are as follows:

1.1 SPSS

SPSS is an abbreviation for Statistical Package for Social Sciences. It is a well-known, proprietary software equipped with graphical user interface. SPSS is especially developed to meet research objectives of social science community, however, research communities across all knowledge domains use the package widely for statistical analysis. The first version of SPSS was released in 1968 for batch processing and mainframes. SPSS is fully owned by IBM Corporation from 2009 onwards. It is available at https://www.ibm.com/in-en/marketplace/statistical-analysis-and-reporting.

1.2 PSPP

PSPP is free alternative statistical package for IBM's SPSS Statistics. It was developed in C using GNU Scientific Library for its mathematical routines as a free alternate to SPSS in 1990s. This software can be downloaded from https://www.gnu.org/software/pspp/get.html.

1.3 Stata

Stata, a platform independent analytical package, was developed in 1985. It is a proprietary and general-purpose integrated statistical package, which was developed by StataCorp. This software is available at http://www.stata.com/.

1.4 SYSTAT

SYSTAT was designed for statistical and graphical presentation of scientific and engineering data. It was developed by Leland Wilkinson, Assistant Professor of Psychology at University of Illinois at Chicago in the late 1970s, which was incorporated in 1983. More details are available at https://systatsoftware.com/.

1.5 SAS

SAS (Statistical Analysis System) is a commercial software developed by SAS Institute. The new statistical methods, additional routines and JMP had been integrated in the 1980s and 1990s. The trial version of SAS Data Loader for Hadoop, trial version of JMP, demo version of SAS Visual Statistics and free version of SAS University Edition - Free Software for Learning are available at https://www.sas.com/ .

1.6 Minitab

Minitab is statistical software developed by researchers namely, F.R. Barbara, A.R. Thomas and L.J. Brian at the Pennsylvania State University in 1972. It provides e-Learning routine for teaching statistical tools and concepts for quality improvement and offers a Six Sigma and Lean Manufacturing management tool. The standard commercial version of Minitab can be obtained at https://www.minitab.com/ .

1.7 Statistica

Statistica is a suite of window-based proprietary software packages for numerical analysis which is originally derived from a set of statistical packages and plug-ins during Mid-1980 by StatSoft. It was acquired by Dell in March 2014. A free trial version of Statistica can be downloaded at http://statistica.io/ .

1.8 SOFA Statistics

SOFA Statistics is an open source, non-proprietary statistical package for statistical and data analysis which is written in Python. The "SOFA" is abbreviation for *Statistics Open For All*. It provides a user-friendly graphical user interface for analysis. It is available at http://www.sofastatistics.com/home.php.

2. R: An Open Source Statistical Package

R is an open source software development programming language, which was developed by Ross Ihaka and Robert Gentleman in New Zealand in 1991. It is a programming language, which can be used in statistical analysis and in technical research work. Before R, John Chambers and his colleagues implemented S language in Bell Laboratory in 1976. R became an open source software in 1997 under GNU General Public License. R is becoming a popular programming language amongst statistician and data miners for the development of statistical packages and analyzing the data.

R is a well-developed programming language, which includes various functionalities and facilities including user-defined recursive functions, conditions, loops and input and output data. R has large collection of operators to perform calculations on different data types like arrays, vectors, lists, and matrices. R is simple and very effective in data handling and storage facility. It has wide range of comprehensible and integrated tools for analyzing large datasets. R is also being taught in various universities.

2.1 Key Features of R

R is more than just a programming language. Object oriented programming can be done using R language. The key features of R are given below.

2.1.1 Packaged Distribution

R software package is a bundle of R functions. Like ".jar" file in Java and ".dll" file in DOT NET, R software package is self-contained units of R functionality and can be used as functions. R software package can be downloaded from CRAN (Comprehensive R Archive Network) repositories. CRAN has country wise repository as well as it has main repository from where software packages (as per requirement) can be downloaded and installed. Vast library of R software packages are available with various operations including machine learning, graphical representation of data, statistical operations, etc. List of R software packages available on CRAN can be found at http://cran.r-project.org.

As R Programming language is open source, one can develop and distribute R software package as per the requirement and can contribute to open source community so as to extend benefit to others.

2.1.2 Data Operations

R programming language enables a wide range of operations. Various operations like Data Cleaning, Exploration and Analysis can be performed using R very easily. Descriptive and predictive analytics with visualization of resultant data can be performed using R. R performs statistical operations like regression, probability distribution, mean, min, max, quartiles, variance and many more. Machine learning operations i.e. linear regression, logistic regression, classification and clustering can be processed using R.

Third-party APIs are also available to develop effective analytics applications. R has the packages to perform various operations on data. Other programming languages like java, c, c++, PHP can also be integrated with R.

2.1.3 Community Support

R is becoming popular programming language for statistical analysis, machine learning and data visualization. Community support for R language is also growing with its popularity. Several R groups and communities are available for users to connect and get help on problems and issues that they may face. R project owners have created official group named "R mailing list". Besides, several blogs are available where user can get the source code. Developer can post their problems over Stack overflow (http://stats.stackexchange.com) where they can get the solution. Lots of books and guides are available for learning R language from beginner to advanced level.

2.1.4 Data Modelling in R

Machine learning is an artificial intelligence (AI) technique that provides systems with the ability to learn without any programme. Data modeling is very important part in machine learning to understand the hidden pattern from the old dataset. This will help in future to predict the similar data. In this technique, the system focuses on user's activities and operations to understand their preferences and tastes. Well known organizations like Amazon, Google, Facebook, LinkedIn, Twitter, etc. have started this data modeling techniques to understand the behaviour of the customer. R has very efficient support for various modeling technique as mentioned below:

2.1.4.1 Regression: This technique is to understand the scalar relationship between variables that will help to predict the value of variables for future events. Forecasting or prediction can be made using this technique. R has lm method to provide regression feature.

2.1.4.2 Classification: Classification is the technique that is used to classify the set of observations into labels. R has various packages available to serve classification, i.e. randomForest, glm, glm2, glmnet, svm and ksvm. Email service providers are using this technique to filter and classify emails as a spam or others.

2.1.4.3 Clustering: Clustering is the technique to group similar items in organized manner from the collection of items. Various packages like knn, kmeans, dist, pvclust and Mclustmethods are available in R to work with clustering.

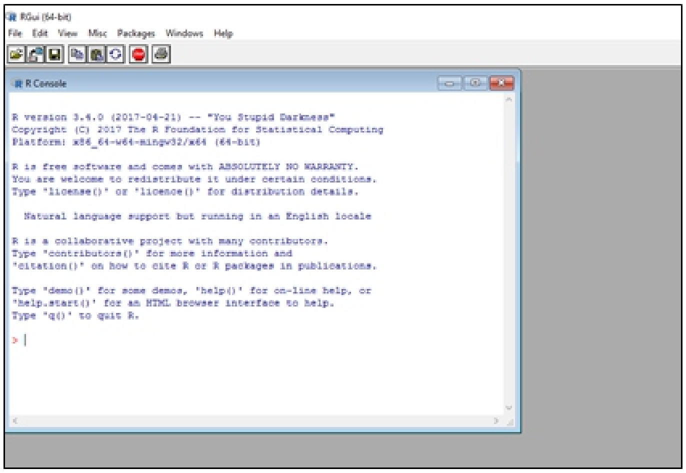
2.1.4.4 Recommendation: This technique is generally applied for assisting and augmenting the process and also help in making decisions. Various recommendation algorithms are being used in recommender systems today. Web content recommendations may include similar websites, blogs, videos, or related content. Recommender package is available in R to use Recommendation technique.

2.2 Integrated Development Environment

R is command line language. It supports many integrated development environments and editors are available for R programming. All the tools are freely available that can be used for statistical analysis or development purpose. Brief details about different tools are given below:

2.2.1 R GUI (Graphical User Interface)

RGui comes as a default and standard software development environment. RGUI is basic user interface. It provides console to write instructions, scripts and general R operations. Console is very light weight, as such, script execution becomes faster comparatively. R Console provides basic information about the installed version of R and brief information about calling functions in R. User can write command/function after the symbol " > " is displayed at the end. Screenshot of RGUI is given below:



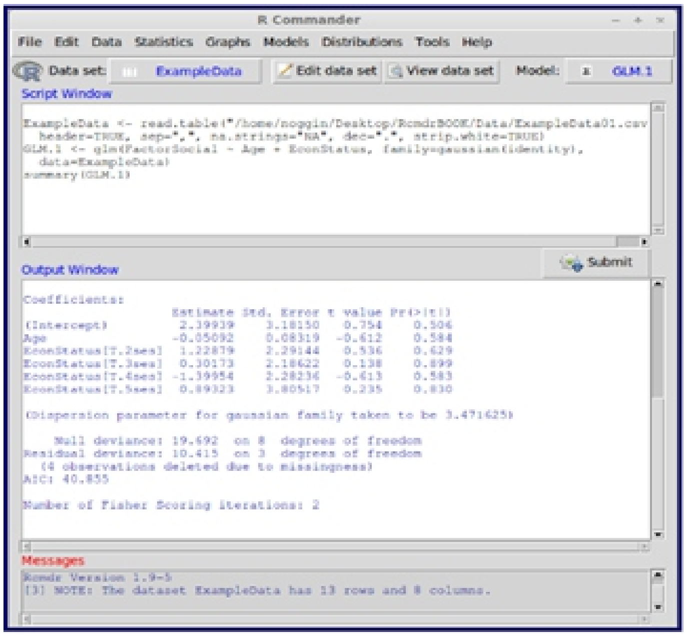
```
R version 3.4.0 (2017-04-21) -- "You Stupid Darkness"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

RGUI

2.2.2 RCommander

RCommander, also known as Rcmdr, is a very powerful tool for data analysis. R Commander is the graphical user interface for analyzing the data using R engine in the backend. RCommander can be installed as a package in R. Novice users can immensely benefit from it as it is displays the underlying R command. It provides lots of packages like analyses, graphics, books and teaching as plugins of RCommander. It can also be used in conjunction with Rstudio.

RCommander provides the point and click interface by which user can access the statistical methods. It displays the background command so that one can also use the same in future if required. Minitab, SPSS, SAS, Stata, Excel and plain-text (ASCII) file can be imported using Rcommander.



```
ExampleData <- read.table("/home/noggin/Desktop/RcmdrBOOK/Data/ExampleData01.csv
  header=TRUE, sep=",", na.strings="NA", dec=".", strip.white=TRUE)
GLM.1 <- glm(FactorSocial ~ Age + EconStatus, family=gaussian(identity),
  data=ExampleData)
summary(GLM.1)
```

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.39939    3.18150   0.754    0.506
Age                -0.05092    0.08319  -0.612    0.584
EconStatus[T.2ses]  1.22879    2.29144   0.536    0.629
EconStatus[T.3ses]  0.30173    2.18622   0.138    0.899
EconStatus[T.4ses] -1.39954    2.28236  -0.613    0.583
EconStatus[T.5ses]  0.89323    3.80517   0.235    0.830

(Dispersion parameter for gaussian family taken to be 3.471625)

    Null deviance: 19.692  on 8  degrees of freedom
Residual deviance: 10.415  on 3  degrees of freedom
  (4 observations deleted due to missingness)
AIC: 40.855

Number of Fisher Scoring iterations: 2
```

```
Rcmdr Version 1.9-5
[3] NOTE: The dataset ExampleData has 13 rows and 8 columns.
```

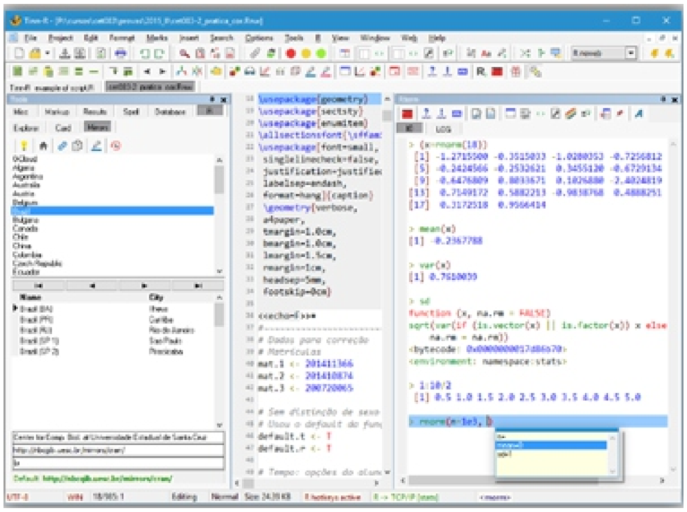
Rcommander

2.2.3 Tinn-R

Tinn-R is an open source project under the GNU General Public License. Tinn-R is coordinated project that aims to facilitate learning and use of R for statistics. It provides advance features for experts like editing, processing, compilation of documents and format interchange using R, Noweb, LaTeX, Pandoc, Txt2tags, etc.

It is an editor/word processor. ASCII / UNICODE, generic for the Windows operating system, is well integrated into the R, with characteristics of Graphical

User Interface (GUI) and Integrated Development Environment (IDE). Tinn-R is stable and have a great structural simplicity with very good performance. Screenshot of Tinn-R is given below:



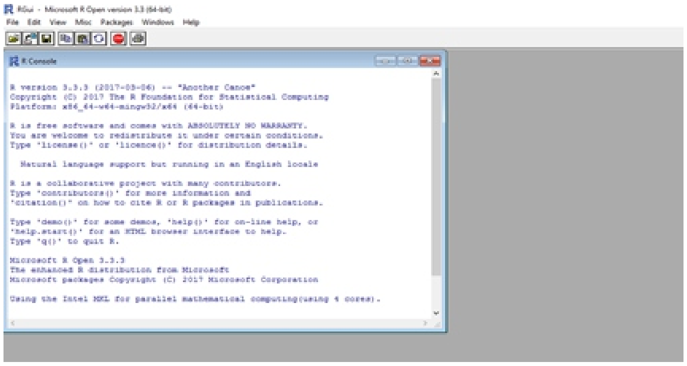
```
\usepackage{geometry}
\usepackage{sectsty}
\usepackage{enumitem}
\allsectionsfont{\sffam
\usepackage[font=small,
 singlelinecheck=false,
 justification=justifie
 labelsep=endash,
 format=hang]{caption}
 \geometry{verbose,
 a4paper,
 tmargin=1.0cm,
 bmargin=1.0cm,
 lmargin=1.5cm,
 rmargin=1cm,
 headsep=5mm,
 footskip=0cm}

<<echo=F>>=
#-------------------------
# Dados para correção
# Matrículas
mat.1 <- 201411366
mat.2 <- 201410874
mat.3 <- 200720065

# Sem distinção de sexo
# Usou o default da func
default.t <- T
default.r <- T

# Tempo: opções do alun
```

```
> (x<-rnorm(18))
 [1] -1.2715500 -0.3515033 -1.0280353 -0.7256812
 [5] -0.2424566 -0.2532621  0.3455120 -0.6720134
 [9] -0.6476809  0.8033671  0.1026880 -2.4024819
[13]  0.7149172  0.5882213 -0.9838768  0.4888251
[17]  0.3172518  0.9566414

> mean(x)
[1] -0.2367788

> var(x)
[1] 0.7610039

> sd
function (x, na.rm = FALSE)
sqrt(var(if (is.vector(x) || is.factor(x)) x else
    na.rm = na.rm))
<bytecode: 0x00000000017d86b70>
<environment: namespace:stats>

> 1:10/2
 [1] 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0

> rnorm(n=1e3, )
```

Tinn-R

2.2.4 Microsoft R Open & R Server

Microsoft Corporation has also published its own tool with integration of R language, which is named as "Microsoft R Open", formerly known as Revolution R Open (RRO). Latest version of Microsoft R Open is 3.4.0 with R-3.4.0. Microsoft R Open is having performance improvement and capabilities for reproducibility and various platform support. Microsoft Corporation has contributed in Open Source movement by making this tool available freely as open source software.



RGui - Microsoft R Open version 3.3 (64-bit)
File Edit View Misc Packages Windows Help
R Console

```
R version 3.3.3 (2017-03-06) -- "Another Canoe"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

Microsoft R Open 3.3.3
The enhanced R distribution from Microsoft
Microsoft packages Copyright (C) 2017 Microsoft Corporation

Using the Intel MKL for parallel mathematical computing(using 4 cores).
```

Microsoft R Open

Microsoft R Open Tool supports Linux platform as well as Windows platform. Main advantage of using this tool is to get the specialized packages created by Microsoft Corporation along with the packages that are available in R. It has multi-threaded math libraries for multi-threaded computations and a high performance CRAN repository for availability of the R packages.

2.2.5 RStudio

RStudio is an integrated development environment (IDE) for R. It is available as Desktop and Server Version. Desktop version of RStudio supports Windows, Linux and Mac. It has open source edition as well as commercial edition. Currently RStudio server supports Linux operating system. RStudio server is also available in commercial version and open source VERSION. RStudio server can be accessed via browser over the Internet. Open source version of RStudio server is sufficient for development purposes. Commercial version of RStudio server has additional features like Administrative Tools, Enhanced Security and Authentication, Metrics and Monitoring and Advanced Resource Management. It is divided in four major parts mentioned below:

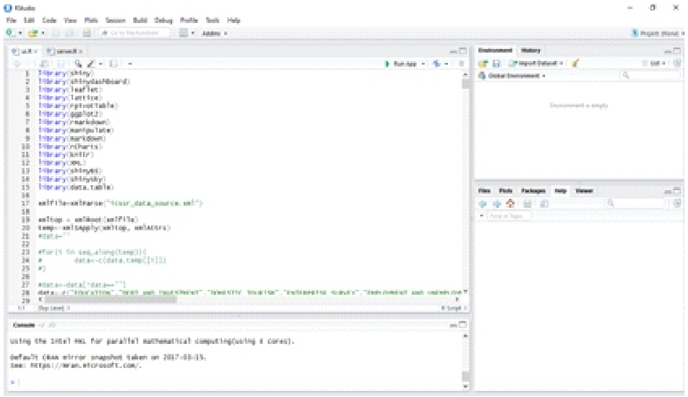
i) Source Editor or Data Viewer

ii) Command History and Workspace Browser

iii) R Console

iv) File, Help, Package and Plots Panel

It is easy to manage more than one working directories using RStudio. It has debugging option to diagnose and fix the errors efficiently. It has integrated documentation



```
library(shiny)
library(shinydashboard)
library(leaflet)
library(lattice)
library(rpivotTable)
library(ggplot2)
library(rmarkdown)
library(manipulate)
library(markdown)
library(rCharts)
library(knitr)
library(XML)
library(shinyBS)
library(shinysky)
library(data.table)

xmlfile=xmlParse("[illegible]_data_source.xml")

xmltop = xmlRoot(xmlfile)
temp<-xmlSApply(xmltop, xmlAttrs)
#data<-""

#for(i in seq_along(temp)){
#        data<-c(data,temp[[i]])
#}

#data<-data[-data==""]
data<-c([illegible]
```

```
Using the Intel MKL for parallel mathematical computing(using 4 cores).

Default CRAN mirror snapshot taken on 2017-03-15.
See: https://mran.microsoft.com/.

>
```

Rstudio

for each function. Navigation to files and functions is very quick. It supports SourceSafe like Git and Subversion. In RScript file, one can use syntax highlighting, code completion and smart indention.

2.2.6 Web Interface using Shiny

Shiny package is used to develop web application without having much knowledge of HTML, CSS or JavaScript. It can be installed using CRAN (Comprehensive R Archive Network). Shiny web application are generally built using R programming and works in any R environment such as Console R, Rgui and RStudio, etc. Shiny Package provides pre-built controls like plots, tables and customizable slider with animation and attractive default UI theme. It uses reactive programming model that manages event handling. It has fast bidirectional communication between browser and R environment.

2.3 Major Advantages of R

2.3.1 Open Source and Free: R is available as open source software licensed under GNU General Public License and it is free to download and use. Thousands of packages are also available under the same license which can be used. Packages can be downloaded from CRAN repositories without any access restrictions.

2.3.2 R is popular - and increasing in popularity: R is becoming popular like other languages even though it is domain-specific language for statistics. This not only shows the increasing interest in R as a programming language, but also of the fields like Data Science and Machine Learning where R is commonly used. IEEE has given rank 5 to R language in 2016. Statisticians are preferring R language for providing quick results of statistical operations. R is becoming popular amongst researchers and post-doctorate students for its reproducible research feature. Data scientists are also using R language for data analytics.

2.3.3 R runs on all platforms: R supports cross-platform interoperability. A such, it can be executed on Windows, Linux as well as Mac. R code written on Windows operating system can run on other operating system.

2.3.4 R is being used by the biggest tech giants/companies: Many tech giants have adopted R language for decision making and data analysis. Every major decision has to be backed by concrete analysis of data. R has the potentiality and power to perform quick analysis and it is too simple. Few industries and platforms that have started using R are listed below:

i) Twitter uses R to monitor user experience

ii) Ford analyses social media to support design decisions of cars using R language

iii) Microsoft Corporation released its R open (an enhanced R distribution) and R server after Revolution Analytics in 2015.

iv) Human Rights Data Analysis Group is using R for measuring the impact of war.

v) Google has also contributed to R by creating R style guide for R community

3. Conclusion

This article provides a brief summary of various statistical packages used for statistical and data analysis, data mining and data visualization in research community. A total number of nine statistical packages including proprietary, open access and open source packages are summarized in this article that are widely used by the academic and scientific research community across the globe. All statistical packages mentioned above have their own set of features, characteristics and functions to perform specific task. It is left to the user to choose a package that is most suitable to them based on their own comfort level, help and expertise available to them.

4. References

i. GNU PSSP available at https://www.gnu.org/s/pspp/ (Accessed on 19/04/2017)

ii. PSPP available at https://en.wikipedia.org/wiki/PSPP (Accessed on 19/04/2017)

iii. Stata: User Guide available at http://www.icssrdataservice.in/files/ICSSR%20Data%20Service-%20Stata%2012.1%20User%20Guide.pdf (Accessed on 19/04/2017)

iv. Stata available at https://en.wikipedia.org/wiki/Stata (Accessed on 19/04/2017)

v. T. Krishnan. SYSTAT: An Overview available at http://www.iasri.res.in/ebook/ebadat/1-Computer%20Usage%20and%20Statistical%20Software%20Packages/9-SYSTAT%20TUTORIAL _03feb.pdf (Accessed on 19/04/2017)

vi. Systat available at http://www.stat.purdue.edu/~jennings/stat582/software/Systat.pdf (Accessed on 19/04/2017)

vii. SYSTAT (software) available at https://en.wikipedia.org/wiki/SYSTAT_(software) (Accessed on 19/04/2017)

viii. SAS University Edition available at https://www.sas.com/en_in/software/university-edition.html (Accessed on 19/04/2017)

ix. SAS (Software) available at https://en.wikipedia.org/wiki/SAS_(software) (Accessed on 19/04/2017)

x. Minitab available at https://en.wikipedia.org/wiki/Minitab (Accessed on 19/04/2017)

xi. Statistica available at https://en.wikipedia.org/wiki/Statistica (Accessed on 19/04/2017)

xii. Statistica Features Overview available at http://www.statsoft.com/Products/STATISTICA-Features (Accessed on 19/04/2017)

xiii. Sofa Statistics : Features Highlights available at http://www.sofastatistics.com/features.php (Accessed on 19/04/2017)

xiv. Sofa Statistics available at https://en.wikipedia.org/wiki/SOFA_Statistics (Accessed on 19/04/2017)

xv. R (programming language) available at https://en.wikipedia.org/wiki/R_(programming_language) (Accessed on 21/04/2017)

xvi. Lecture-7: An Overview and History of R available at https://www.coursera.org/learn/r-programming/lecture/pAbaE/overview-and-history-of-r (Accessed on 25/04/2017)

xvii. How to navigate R GUI available at http://www.dummies.com/programming/r/how-to-navigate-rgui/ (Accessed on 25/04/2017)

xviii. R commander (Rcmdr): A graphical interface for R available at http://www.rcommander.com/ (Accessed on 25/04/2017)

xix. Tinn-R Editor - GUI for R Language and Environment available at https://sourceforge.net/projects/tinn-r/ (Accessed on 25/04/2017)

xx. Microsoft R Open: The Enhanced R Distribution available at https://mran.microsoft.com/open/ (Accessed on 25/04/2017)

xxi. RStudio: Open Source and Enterprise Ready Professional Software for R available at https://www.rstudio.com/ (Accessed on 25/04/2017)

xxii. Understanding the features of R language available at https://www.packtpub.com/mapt/book/big-data-and-business-intelligence/9781782163282/1/ch01lvl1sec17/understanding-the-features-of-r-language (Accessed on 28/04/2017)

xxiii. Introducing Shiny available at http://rstudio.github.io/shiny/tutorial/ (Accessed on 01/05/2017