Assignment 5
Study of open source analytical software

AIM : Study of platform and Download the open source software (WEKA, R and Python). Document the distinct features and functionality of the software platform.

OBJECTIVE :

To study

· Concept of open source analytical software. (WEKA,R and Python)

· Concept of statistical analysis.

· Distinct features and functionality of open source software

THEORY:

Introduction of WEKA

Weka is open source software under the GNU General Public License.The system is written using object oriented language Java. There are several different levels at which Weka can be used. Weka provides implementations of state-of-the-art data mining and machine learning algorithms. Weka contains modules for data preprocessing, classification, clustering and association rule extraction.

Introduction of R

R is a programming language and software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. R is an implementation of the S programming language combined with lexical scoping semantics inspired by Scheme. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team, of which Chambers is a member. The source code for the R software environment is written primarily in C, Fortran, and R.R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems. R uses a command line interface; there are also several graphical front-ends for it.

Introduction of Python

Python is an interpreted,high-level and general purpose programming language created by Guido van Rossum and first released in 1991 Python’s design philosophy emphasizes code readability with its notable use of significant whitespace.Its language constructs and object oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

Concept:

WEKA:

- **Steps to download and configure the WEKA:**

  Download Weka (the stable version) from http://www.cs.waikato.ac.nz/ml/weka/

  - Choose a self-extracting executable (including Java VM)

  After download is completed, run the self extracting file to install Weka, and use the default set-ups.

- **Features of WEKA:**

  Main features of Weka include:

  - 49 data preprocessing tools
  - 76 classification/regression algorithms
  - 8 clustering algorithms
  - 15 attribute/subset evaluators + 10 search algorithms for feature selection.
  - 3 algorithms for finding association rules
  - 3 graphical user interfaces
  - “The Explorer” (exploratory data analysis)
  - “The Experimenter” (experimental environment)
  - “The KnowledgeFlow” (new process model inspired interface)

- **Hardware or software required**:

  Hardware:

  - 4GB RAM

  Software:

  - Java

  - 64-bit / 32-bits versions of Windows.

  - 64-bit / 32-bits Linux

- **Application of WEKA**:

  The WEKA system has been applied successfully in a variety of areas including the areas of agriculture, machine learning research and education.

- **Limitations of WEKA**:

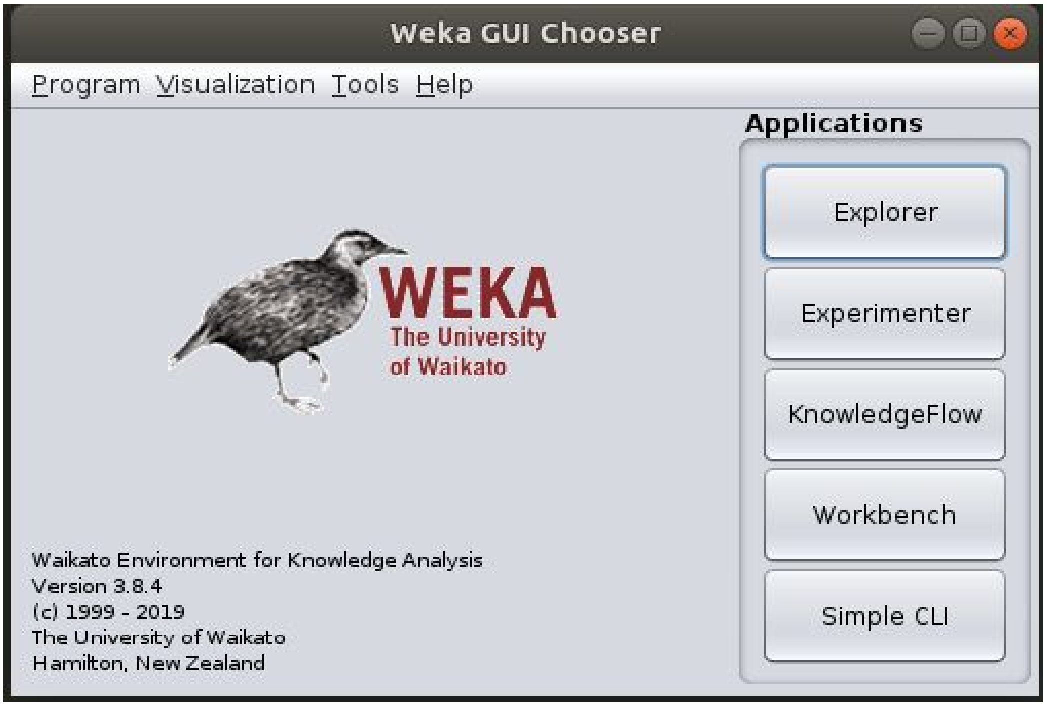
  - GUI is not as well documented.

  -2 different Modules cannot be combined (ex. modules for both PCA and clustering without writing a Java Code).

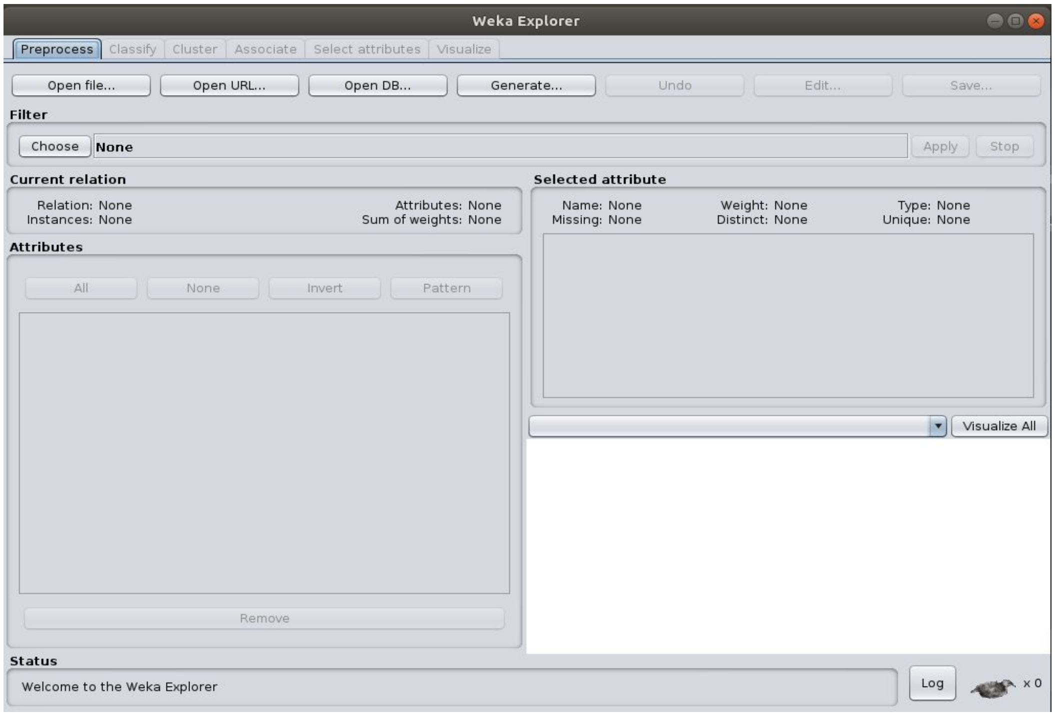
  -The Weka GUI provides several built-in 'visualization' panels but these are very limited.

  -Manipulation of data sets is not easy in Weka

- **Output:**



Waikato Environment for Knowledge Analysis
Version 3.8.4
(c) 1999 - 2019
The University of Waikato
Hamilton, New Zealand

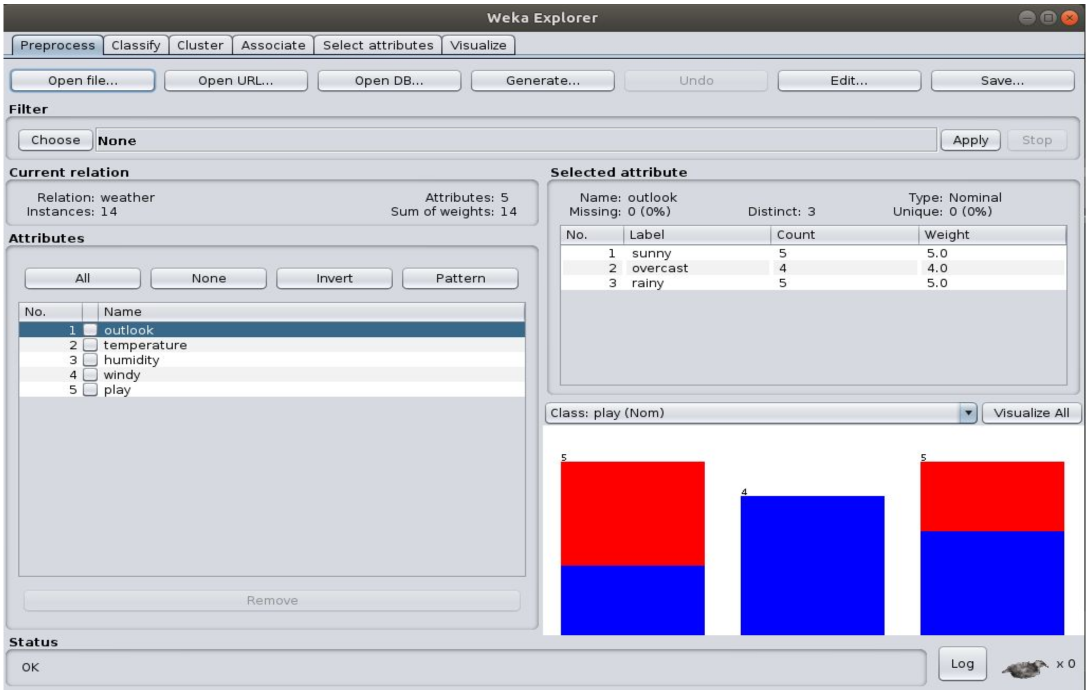


None

Relation: None
Instances: None
Attributes: None
Sum of weights: None

Name: None
Missing: None
Weight: None
Distinct: None
Type: None
Unique: None

Welcome to the Weka Explorer



Filter

None

Current relation

Relation: weather
Instances: 14
Attributes: 5
Sum of weights: 14

Attributes

No.		Name
1	□	outlook
2	□	temperature
3	□	humidity
4	□	windy
5	□	play

Selected attribute

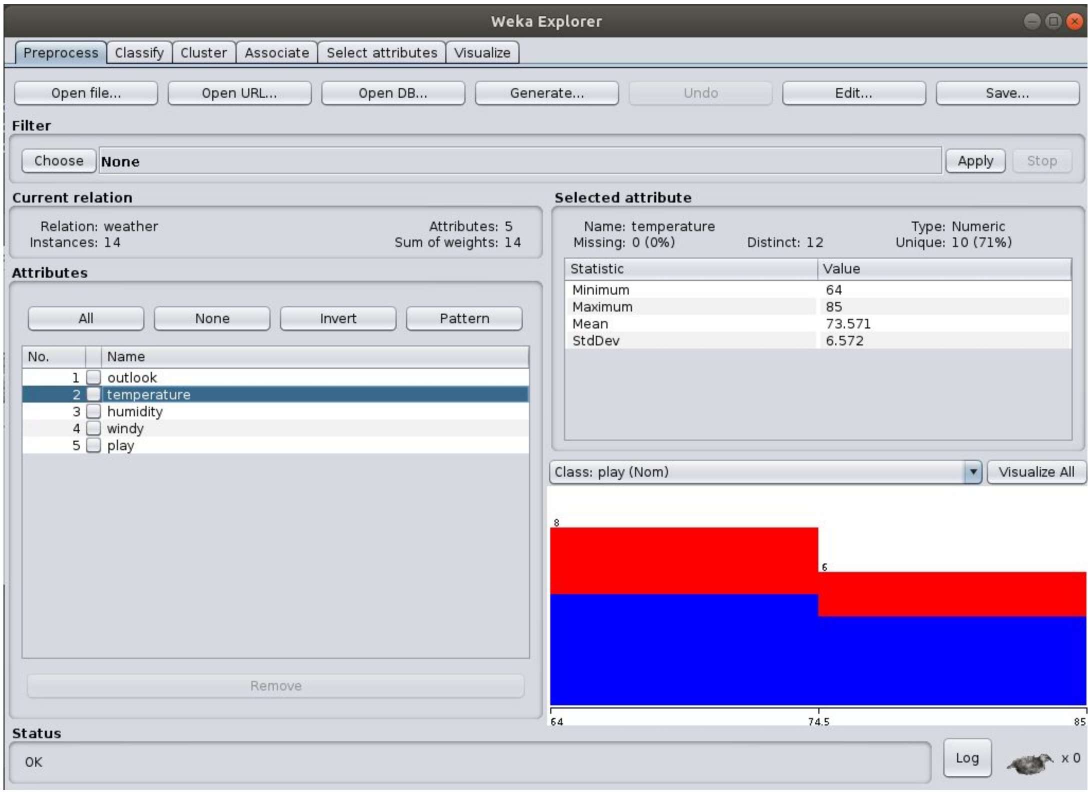
Name: outlook
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	sunny	5	5.0
2	overcast	4	4.0
3	rainy	5	5.0

Class: play (Nom)

Status

OK



Filter

None

Current relation

Relation: weather
Instances: 14
Attributes: 5
Sum of weights: 14

Attributes

No.		Name
1	□	outlook
2	□	temperature
3	□	humidity
4	□	windy
5	□	play

Selected attribute

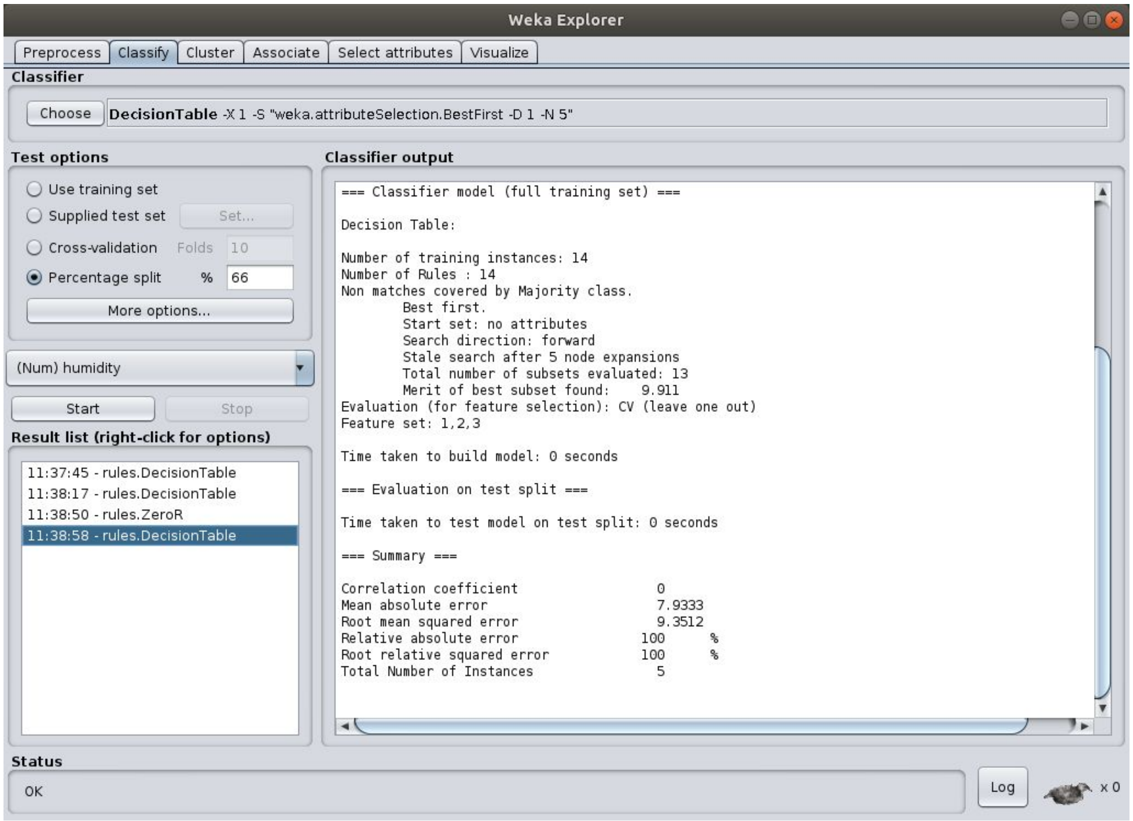
Name: temperature
Missing: 0 (0%)
Distinct: 12
Type: Numeric
Unique: 10 (71%)

Statistic	Value
Minimum	64
Maximum	85
Mean	73.571
StdDev	6.572

Class: play (Nom)

Status

OK



Classifier

DecisionTable -X 1 -S "weka.attributeSelection.BestFirst -D 1 -N 5"

Test options

○ Use training set

○ Supplied test set

○ Cross-validation Folds 10

● Percentage split % 66

(Num) humidity

Result list (right-click for options)

11:37:45 - rules.DecisionTable
11:38:17 - rules.DecisionTable
11:38:50 - rules.ZeroR
11:38:58 - rules.DecisionTable

Classifier output

```
=== Classifier model (full training set) ===

Decision Table:

Number of training instances: 14
Number of Rules : 14
Non matches covered by Majority class.
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 13
        Merit of best subset found:    9.911
Evaluation (for feature selection): CV (leave one out)
Feature set: 1,2,3

Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correlation coefficient                  0
Mean absolute error                      7.9333
Root mean squared error                  9.3512
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances                5
```

Status

OK

R:

- **Steps to download and configure the R:**

  1. Install r-base : Write this command in Command Prompt : sudo apt-get install r-base

  2. Type R on terminal/Command line to get command line for R programming.

  Using IDE:

  1. URL for Rstudio :

  http://www.rstudio.com/products/rstudio/download/

  Write this command in Command Prompt: sudo apt-get install r-base

  OR

  2. Open Ubuntu Software Center

  Search : R-studio

Install : R-Studio

- **Features of R:**

  Statistical Features and Implement a wide variety of statistical and graphical techniques.

  - R is easily extensible through functions and extensions, and the R community is noted for its active contributions in terms of packages.

  - For computationally intensive tasks, C, C++, and Fortran code can be linked and called at run time.

  Programming Features

  - R is an interpreted language.

  -R's data structures include vectors, matrices, arrays, data frames (similar to tables in a relational database) and lists.

  - R supports procedural programming with functions and, for some functions, object-oriented programming with generic functions.

- **Hardware or Software required:**

  Hardware:

  - The amount of RAM that you need is highly dependent on the work/analysis you will be doing. (More than 1 GB of RAM.)

  Software:

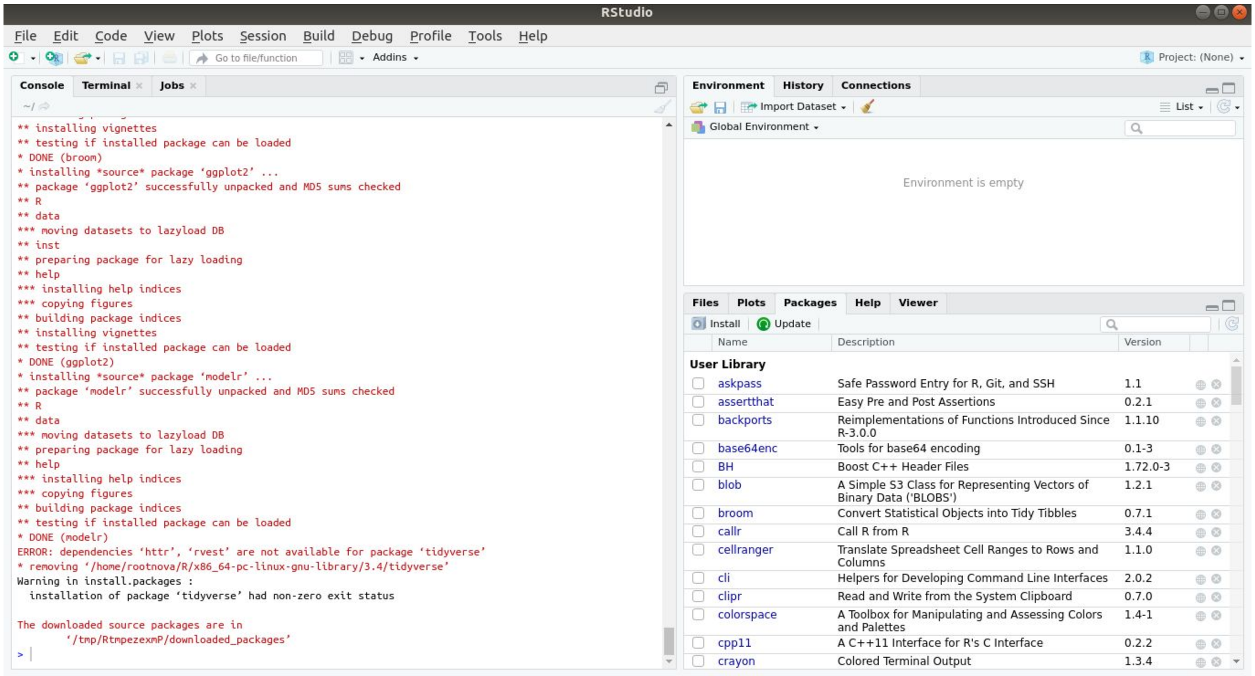
  - 64-bit / 32-bits versions of Windows.

  - 64-bit / 32-bits Linux

- **Application of R:**

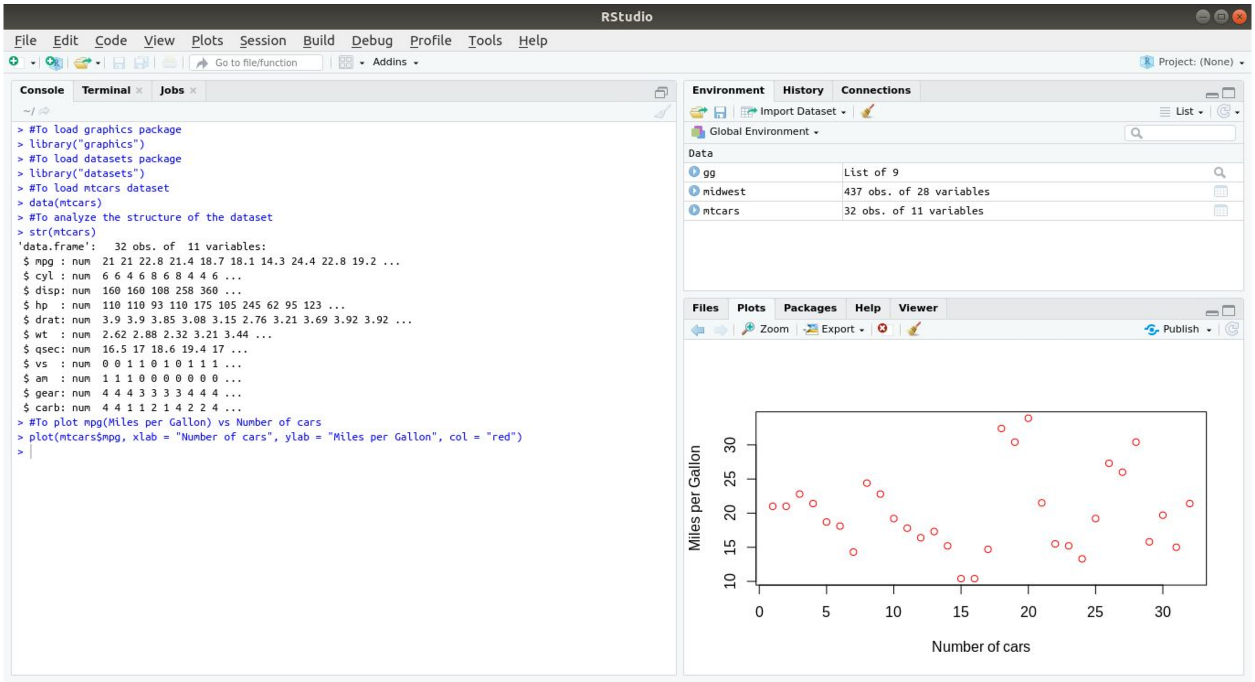
  R applications span the universe from theoretical computational statistics and the hard sciences such as astronomy, chemistry and genomics to practical applications in business, drug development, finance, health care, marketing, medicine and much more.

- **Output:**



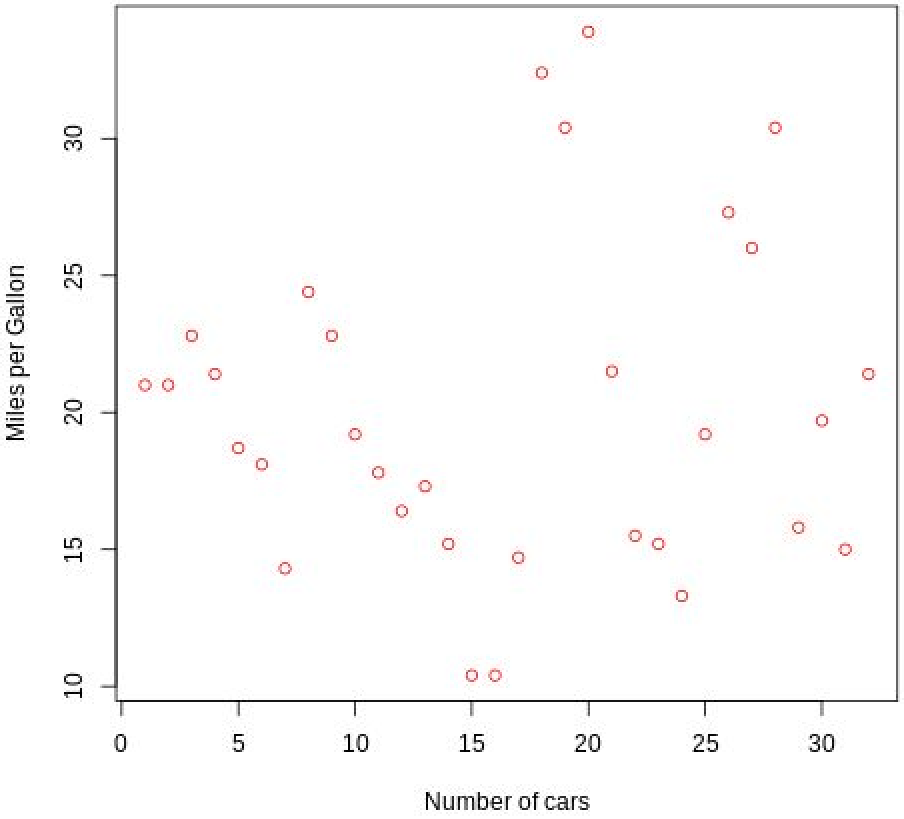
```
** installing vignettes
** testing if installed package can be loaded
* DONE (broom)
* installing *source* package ‘ggplot2’ ...
** package ‘ggplot2’ successfully unpacked and MD5 sums checked
** R
** data
*** moving datasets to lazyload DB
** inst
** preparing package for lazy loading
** help
*** installing help indices
*** copying figures
** building package indices
** installing vignettes
** testing if installed package can be loaded
* DONE (ggplot2)
* installing *source* package ‘modelr’ ...
** package ‘modelr’ successfully unpacked and MD5 sums checked
** R
** data
*** moving datasets to lazyload DB
** preparing package for lazy loading
** help
*** installing help indices
*** copying figures
** building package indices
** testing if installed package can be loaded
* DONE (modelr)
ERROR: dependencies ‘httr’, ‘rvest’ are not available for package ‘tidyverse’
* removing ‘/home/rootnova/R/x86_64-pc-linux-gnu-library/3.4/tidyverse’
Warning in install.packages :
  installation of package ‘tidyverse’ had non-zero exit status

The downloaded source packages are in
        ‘/tmp/RtmpezexmP/downloaded_packages’
> 
```

	Name	Description	Version
	User Library		
□	askpass	Safe Password Entry for R, Git, and SSH	1.1
□	assertthat	Easy Pre and Post Assertions	0.2.1
□	backports	Reimplementations of Functions Introduced Since R-3.0.0	1.1.10
□	base64enc	Tools for base64 encoding	0.1-3
□	BH	Boost C++ Header Files	1.72.0-3
□	blob	A Simple S3 Class for Representing Vectors of Binary Data ('BLOBS')	1.2.1
□	broom	Convert Statistical Objects into Tidy Tibbles	0.7.1
□	callr	Call R from R	3.4.4
□	cellranger	Translate Spreadsheet Cell Ranges to Rows and Columns	1.1.0
□	cli	Helpers for Developing Command Line Interfaces	2.0.2
□	clipr	Read and Write from the System Clipboard	0.7.0
□	colorspace	A Toolbox for Manipulating and Assessing Colors and Palettes	1.4-1
□	cpp11	A C++11 Interface for R's C Interface	0.2.2
□	crayon	Colored Terminal Output	1.3.4



```
> #To load graphics package
> library("graphics")
> #To load datasets package
> library("datasets")
> #To load mtcars dataset
> data(mtcars)
> #To analyze the structure of the dataset
> str(mtcars)
'data.frame':   32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
> #To plot mpg(Miles per Gallon) vs Number of cars
> plot(mtcars$mpg, xlab = "Number of cars", ylab = "Miles per Gallon", col = "red")
> 
```

Data	
gg	List of 9
midwest	437 obs. of 28 variables
mtcars	32 obs. of 11 variables


Miles per Gallon
10
15
20
25
30
0
5
10
15
20
25
30
Number of cars

Python:

- **Steps to download and configure the Python:**

  -Update and Refresh Repository Lists with (sudo apt update)

  -Install with (sudo apt install python)

  -Check Python version with (python --version)

  -Run python shell by typing (python) on the terminal

- **Features of Python:**

  -Open Source and Free.

  -Support for GUI.

  -Object Oriented Approach.

  -Highly Portable.

  -Highly Dynamic.

  -High -Level Language.

- **Application of Python:**

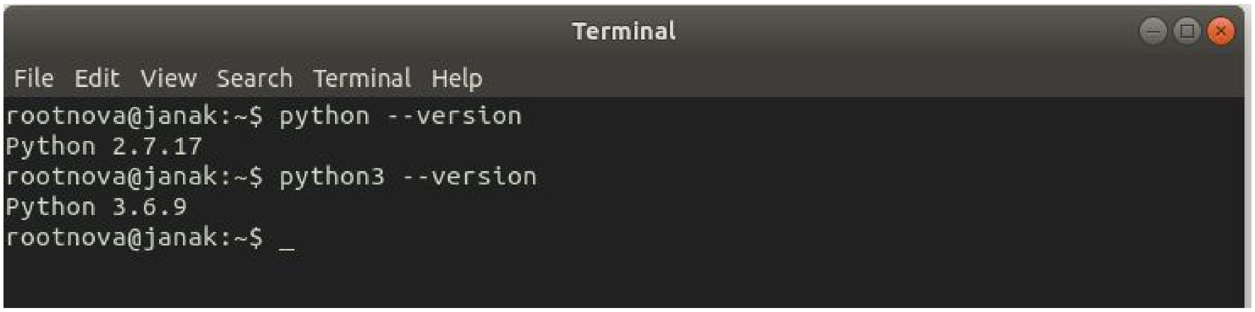
  -Web Applications-It provides libraries to handle internet protocols such as HTML and XML,JSON,Email processing,request,beautifulSoup etc.

  -**Desktop GUI-**Python provides a Tk GUI library to develop a user interface.

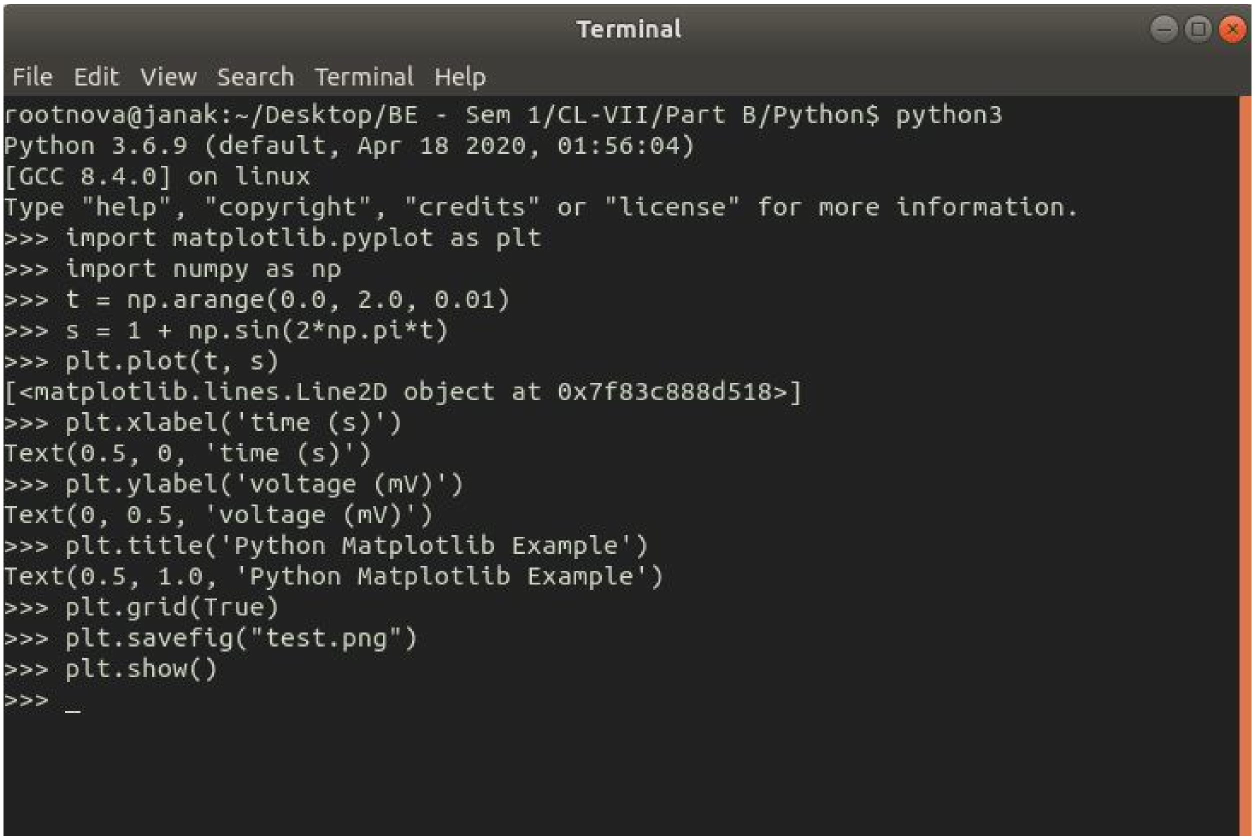
  -**Image Processing-**Python contains many libraries that are used to work with the image. The image can be manipulated according to our requirements. Some libraries of image processing are OpenCV and Pillow.

  -**Business Application-**Business Applications differ from standard applications. E-commerce and ERP are an example of a business application. This kind of application requires extensive scalability and readability, and Python provides all these features.

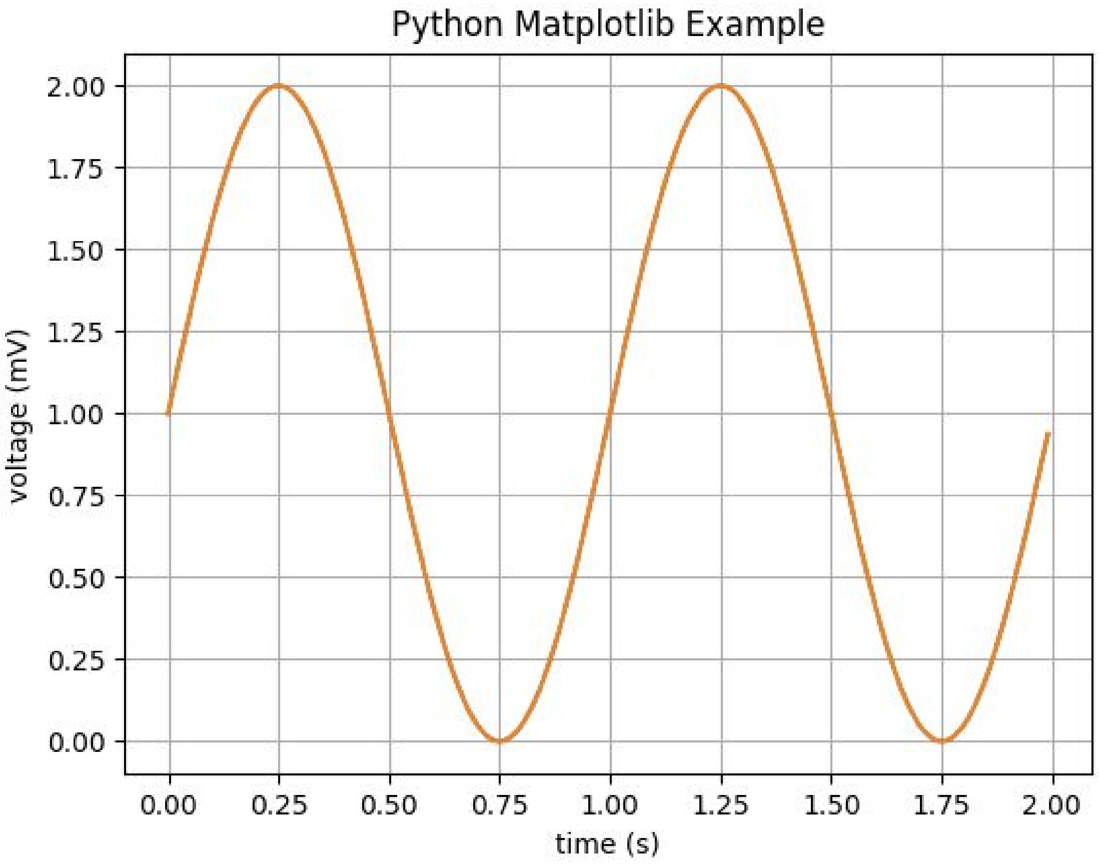
- **Output:**



```
rootnova@janak:~$ python --version
Python 2.7.17
rootnova@janak:~$ python3 --version
Python 3.6.9
rootnova@janak:~$ _
```



```
rootnova@janak:~/Desktop/BE - Sem 1/CL-VII/Part B/Python$ python3
Python 3.6.9 (default, Apr 18 2020, 01:56:04)
[GCC 8.4.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> import matplotlib.pyplot as plt
>>> import numpy as np
>>> t = np.arange(0.0, 2.0, 0.01)
>>> s = 1 + np.sin(2*np.pi*t)
>>> plt.plot(t, s)
[<matplotlib.lines.Line2D object at 0x7f83c888d518>]
>>> plt.xlabel('time (s)')
Text(0.5, 0, 'time (s)')
>>> plt.ylabel('voltage (mV)')
Text(0, 0.5, 'voltage (mV)')
>>> plt.title('Python Matplotlib Example')
Text(0.5, 1.0, 'Python Matplotlib Example')
>>> plt.grid(True)
>>> plt.savefig("test.png")
>>> plt.show()
>>> _
```



CONCLUSION:

Downloaded the open source softwares Python, RStudio and WEKA. Studied the distinct features and functionality of these software platforms.