**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 & 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. System is developed at the University of Waikato in New Zealand. “Weka” stands for the Waikato Environment for Knowledge Analysis. The software is freely available at http://www.cs.waikato.ac.nz/ml/weka. 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 pre-processing, classification, clustering and association rule extraction.

* **Introduction of R**

**R** is a [programming language](https://en.wikipedia.org/wiki/Programming_language) and software environment for [statistical computing](https://en.wikipedia.org/wiki/Statistical_computing) and graphics. The R language is widely used among [statisticians](https://en.wikipedia.org/wiki/Statistician) and [data miners](https://en.wikipedia.org/wiki/Data_mining) for developing [statistical software](https://en.wikipedia.org/wiki/Statistical_software) and data analysis. Polls, [surveys of data miners](https://en.wikipedia.org/wiki/Rexer's_Annual_Data_Miner_Survey), and studies of scholarly literature databases show that R's popularity has increased substantially in recent years. R is an implementation of the [S programming language](https://en.wikipedia.org/wiki/S_(programming_language)) combined with lexical  [scoping](https://en.wikipedia.org/wiki/Lexical_scoping) semantics inspired by [Scheme](https://en.wikipedia.org/wiki/Scheme_(programming_language)). [S](https://en.wikipedia.org/wiki/S_(programming_language)) was created by [John Chambers](https://en.wikipedia.org/wiki/John_Chambers_(programmer)) while at [Bell Labs](https://en.wikipedia.org/wiki/Bell_Laboratories). There are some important differences, but much of the code written for S runs unaltered. R was created by [Ross Ihaka](https://en.wikipedia.org/wiki/Ross_Ihaka) and [Robert Gentleman](https://en.wikipedia.org/wiki/Robert_Gentleman_(statistician)) at the [University of Auckland](https://en.wikipedia.org/wiki/University_of_Auckland), New Zealand, and is currently developed by the *R Development Core Team*, of which Chambers is a member. R is named partly after the first names of the first two R authors and partly as a play on the name of [S](https://en.wikipedia.org/wiki/S_(programming_language)).R is a [GNU project](https://en.wikipedia.org/wiki/GNU_project). The [source code](https://en.wikipedia.org/wiki/Source_code) for the R software environment is written primarily in [C](https://en.wikipedia.org/wiki/C_(programming_language)), [Fortran](https://en.wikipedia.org/wiki/Fortran), and R.R is freely available under the [GNU General Public License](https://en.wikipedia.org/wiki/GNU_General_Public_License), and pre-compiled binary versions are provided for various [operating systems](https://en.wikipedia.org/wiki/Operating_system). R uses a [command line interface](https://en.wikipedia.org/wiki/Command_line_interface); there are also several [graphical front-ends](https://en.wikipedia.org/wiki/Graphical_user_interface) for it.

* **Introduction to 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.

**Concept/Working:**

**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)

– (If you are interested in modifying/extending Weka there is a developer

version that includes the source code)

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

the default set-ups.

* **Working of WEKA**

The general working steps are given below by considering the example of Hierarchical clustering.

1. Select a dataset for example iris.arff.

2. Select option Cluster

3. Choose cluster type: Hierarchical Cluster

4. Select cluster mode: Training Set.

5. Click on Start.

* **Features of WEKA**

Main features of Weka include:

* 49 data pre-processing 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 Knowledge Flow” (new process model inspired interface)

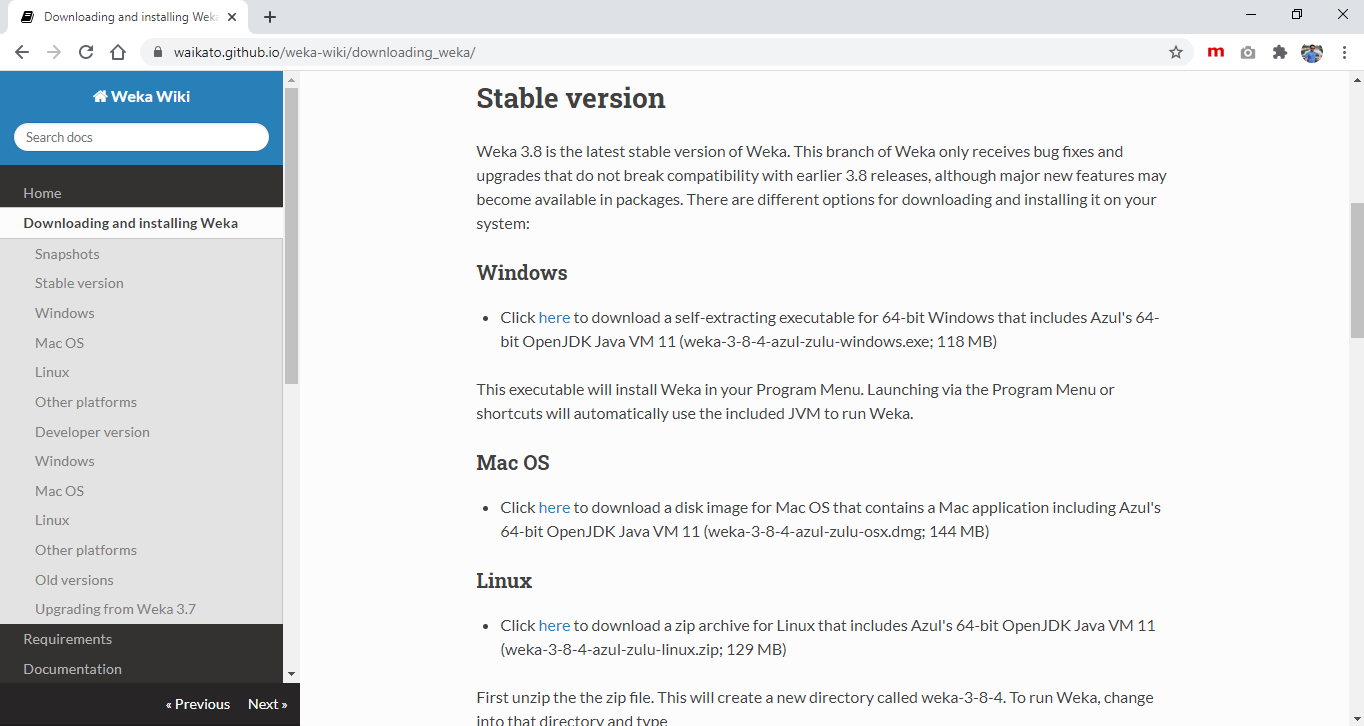
* **Hardware or software required for WEKA.**

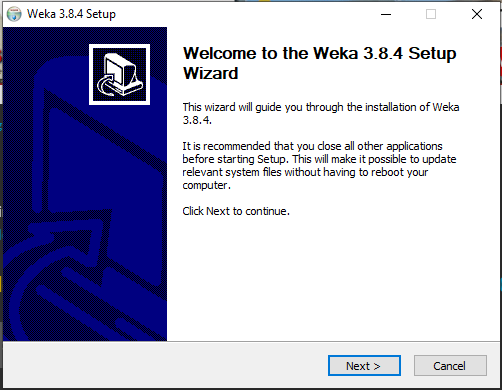
Hardware:

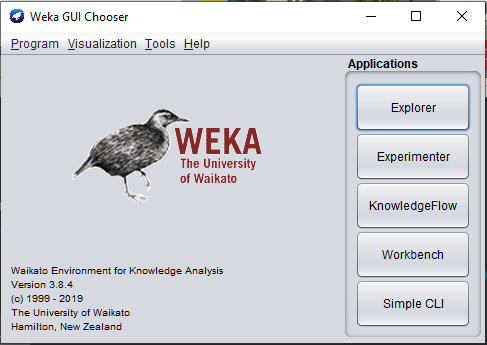
* 4GB RAM

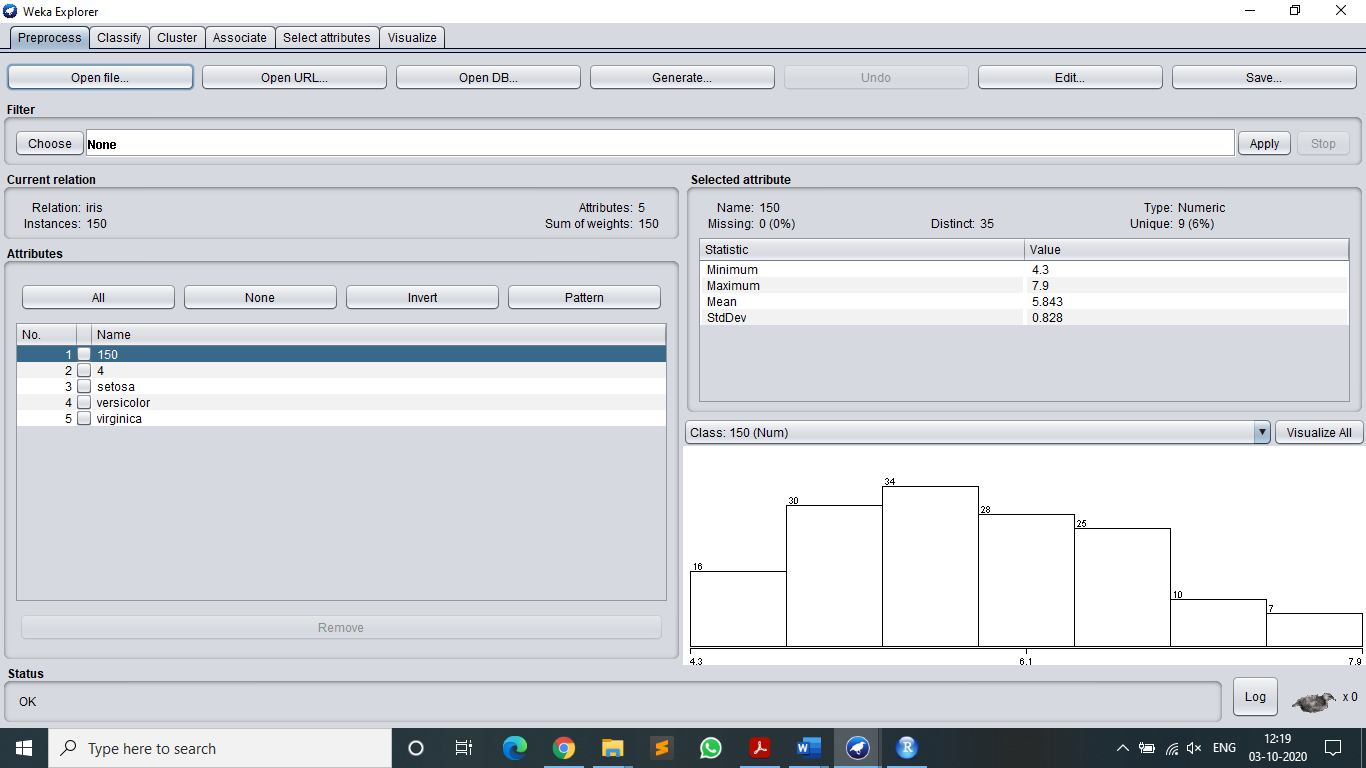
Software:

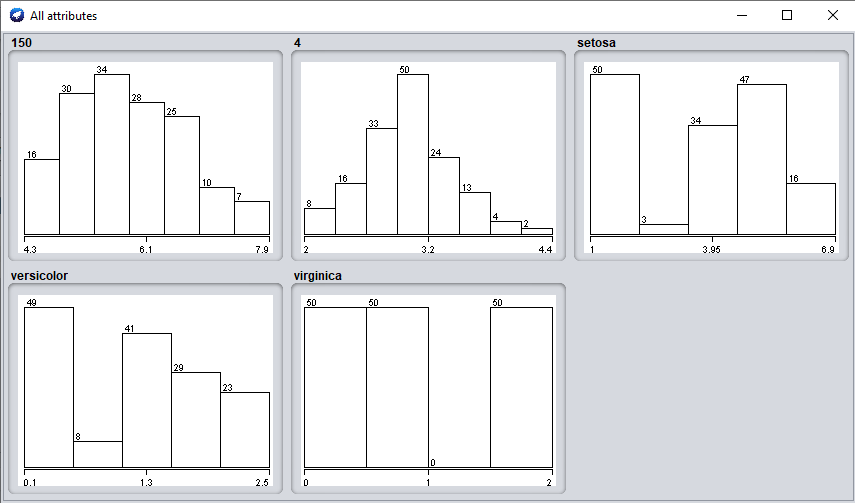
* Java
* 64-bit / 32-bits versions of Windows.
* 64-bit / 32-bits Linux
* **Screenshots**

****

****

****

****

****

* **Advantages of WEKA.**

Free availability:

* Under the GNU General Public License
  + Portability
  + Fully implemented in the Java programming language and thus runs
* on almost any modern computing platforms
* Windows, Mac OS X and Linux
* Comprehensive collection of data preprocessing and modeling
* techniques
  + Supports standard data mining tasks: data preprocessing, clustering,
* classification, regression, visualization, and feature selection.
  + Easy to use GUI
  + Provides access to SQL databases
  + Using Java Database Connectivity and can process the result
* returned by a database query.
* The obvious advantage of a package like Weka is that a whole range of data preparation, feature selection and data mining algorithms are integrated. This means that only one data format is needed, and trying out and comparing different approaches becomes really easy. The package also comes with a GUI, which should make it easier to use.
* **Disadvantages of WEKA.**
* Sequence modeling is not covered by the algorithms included in the Weka distribution.
* Not capable of multi-relational data mining.
* Memory bound.
* Do not implement the newest techniques. For example, the MLP implemented has a very basic training algorithm (backprop with momentum), and the SVM only uses polynomial kernels, and does not support numeric estimation. Therefore, it will be necessary to combine WEKA with some of the other tools like Netlab or SVM\_torch.
* Though the software is for free: the documentation for the GUI is quite limited.
* Limited scaling. For difficult tasks on large datasets, the running time can become quite long, and java sometimes gives an OutOfMemory error. This problem can be reduced by using the '-mxx' option when calling java, where x is memory size (e.g. '50m'). For large datasets it will always be necessary to reduce the size to be able to work within reasonable time limits.
* The GUI does not implement all the possible options. Things that could be very useful, like scoring of a test set, are not provided in the GUI, but can be called from the command line interface. So sometimes it will be necessary to switch between GUI and command line.
* **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.

**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

1. Open Ubuntu Software Center

Search: R-studio

Install: R-Studio

* **Working of R**

The default panes:

* Console (entire left)
* Workspace/History (tabbed in upper right)
* Files/Plots/Packages/Help (tabbed in lower right)

1. Download package needed for that program.
2. Open RStudio.
3. Install the Required Package. Go to Packages - Click on "Install".

You will get "Install Packages" Window.

Install from: Package Archive File(tar.gz)

Package archive: (Browse the path where you have stored the Package\_name.tar.gz) and Click Install.

1. In Packages - Tick the Package\_name in user library. Click on Update. Click on OK.
2. Write Following Code in R Script (File - New File - R Script) and Save it as "Program1.R"

library(e1071)

data(iris)

set.seed(123)

cm1 <- cmeans(iris[,1:4],10)

bc1 <- bclust(iris[,1:4],3,base.centers=20,iter.base=50,base.method="cmeans")

1. Click on run Icon.

7. Write following commands on 'console' which is in 'RStudio'.

> data("iris")

> library(class)

> library(e1071)

> pairs(iris[1:4], main = "Iris Data(red=setosa,green=versicolor,blue=virginica)",pch= 21, bg = c("red","green3","blue")[unclass (iris$Species)])

> data(iris)

> summary(iris)

> classifier<-naiveBayes(iris[,1:4], iris[,5])

> table(predict(classifier, iris[,-5]), iris[,5])

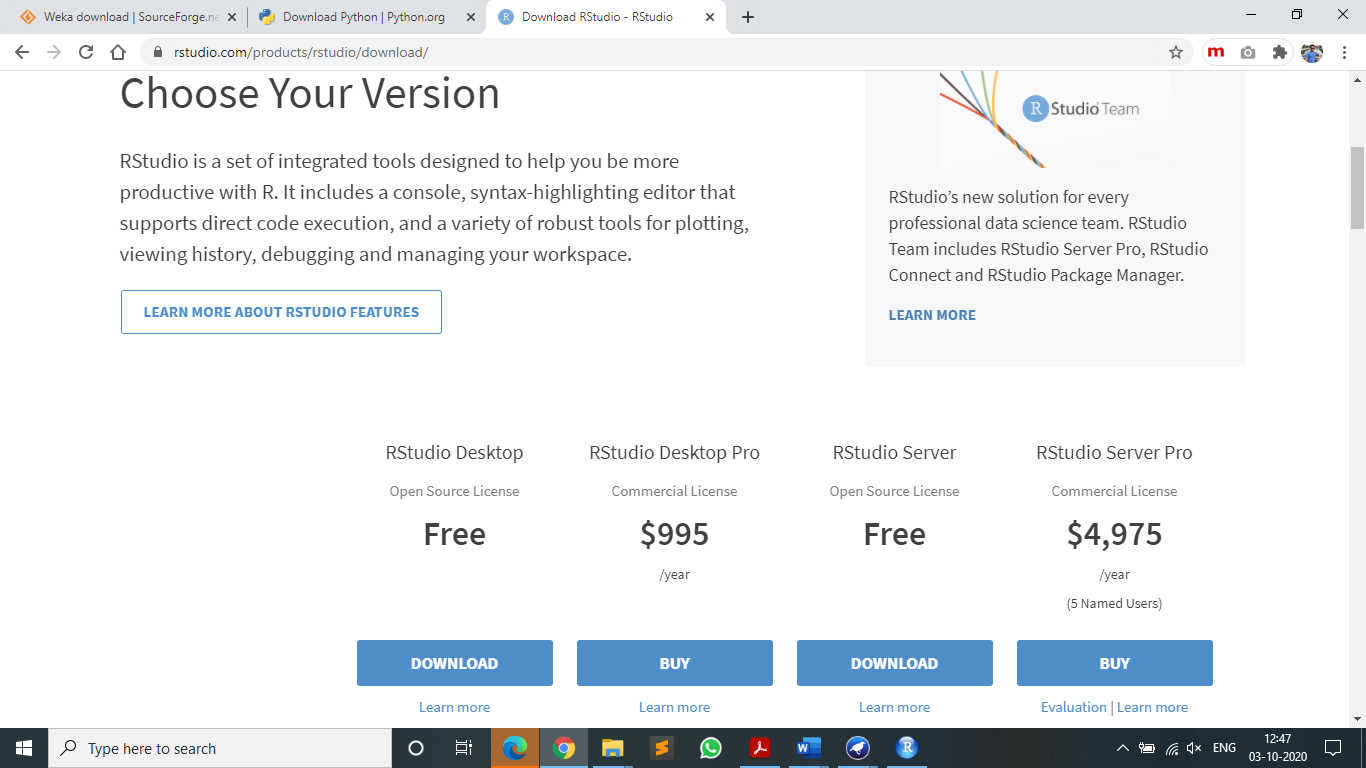
* **Features of R**

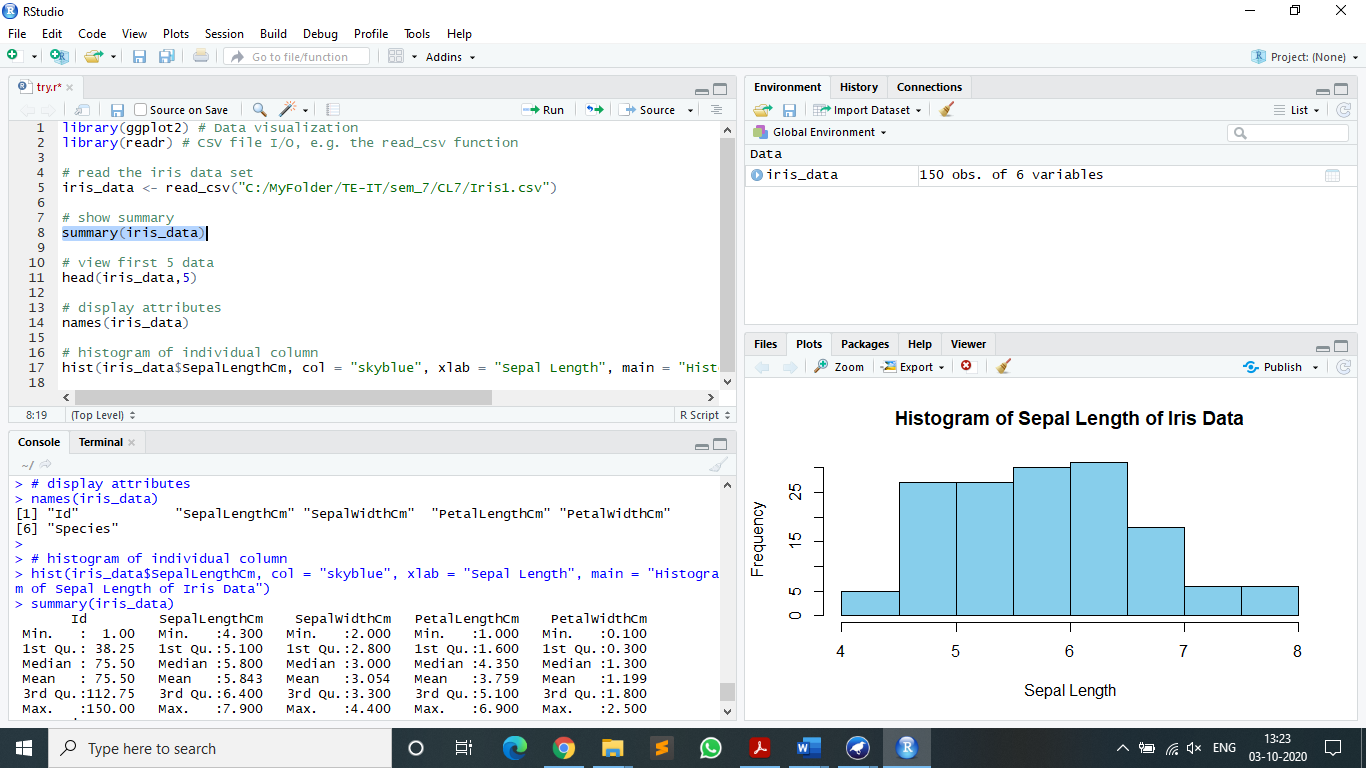
**Statistical Features**

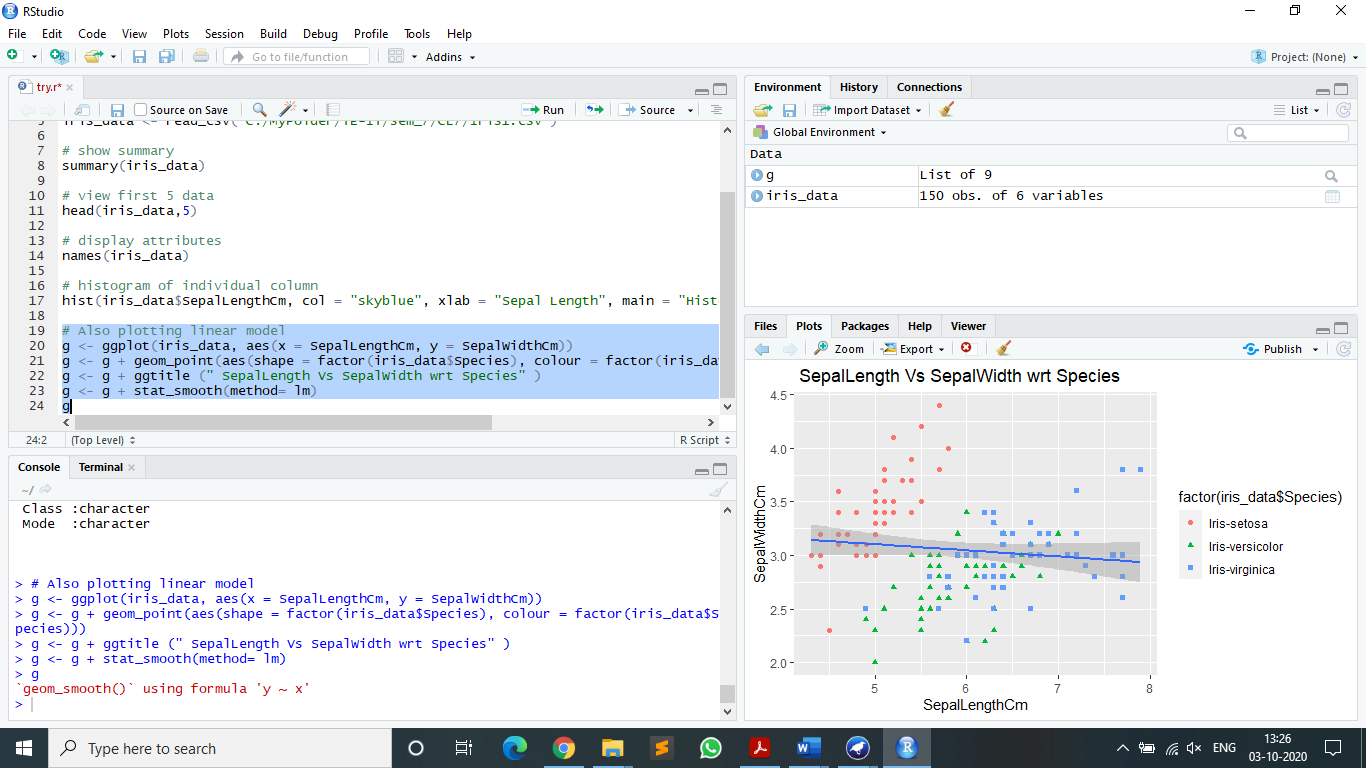
* Implement a wide variety of statistical and [graphical](https://en.wikipedia.org/wiki/Graphical) techniques.
* R is easily extensible through functions and extensions, and the R community is noted for its active contributions in terms of packages.
* For computationally intensive tasks, [C](https://en.wikipedia.org/wiki/C_(programming_language)), [C++](https://en.wikipedia.org/wiki/C%2B%2B), and [Fortran](https://en.wikipedia.org/wiki/Fortran) code can be linked and called at run time.

**Programming Features**

* R is an [interpreted language](https://en.wikipedia.org/wiki/Interpreted_language).
* R's [data structures](https://en.wikipedia.org/wiki/Data_structure) include [vectors](https://en.wikipedia.org/wiki/Column_vector), [matrices](https://en.wikipedia.org/wiki/Matrix_(mathematics)), arrays, data frames (similar to [tables](https://en.wikipedia.org/wiki/Table_(database)) in a [relational database](https://en.wikipedia.org/wiki/Relational_database)) and [lists](https://en.wikipedia.org/wiki/List_(computing)).
* R supports [procedural programming](https://en.wikipedia.org/wiki/Procedural_programming) with [functions](https://en.wikipedia.org/wiki/Function_(computer_science)) and, for some functions, [object-oriented programming](https://en.wikipedia.org/wiki/Object-oriented_programming) with [generic functions](https://en.wikipedia.org/wiki/Generic_function).
* **Screenshots**

****





* **Advantages of R.**
* R is free and open source software, allowing anyone to use and, importantly, to modify it. R is licensed under the GNU General Public License, with copyright held by The R Foundation for Statistical Computing.
* The graphical capabilities of R are outstanding, providing a fully programmable graphics language that surpasses most other statistical and graphical packages.
* R is a programming language and environment developed for statistical analysis by practicing statisticians and researchers.
* **Disadvantages of R.**
* R is not so easy to use for the novice.
* Many R commands give little thought to memory management, and so R can very quickly consume all available memory.
* The quality of some packages is less than perfect.
* **Hardware or Software required for R.**

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. Because R has nearly 5,000 packages (libraries of functions) many of which are dedicated to specific applications you don’t have to be an R genius to begin developing your own applications.

* **Limitations of R**

The biggest limitation in R is the data processing model which is to load everything up in memory and process it. This not only limits the amount of data you can process but it also scales very badly for complex processes.

**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.

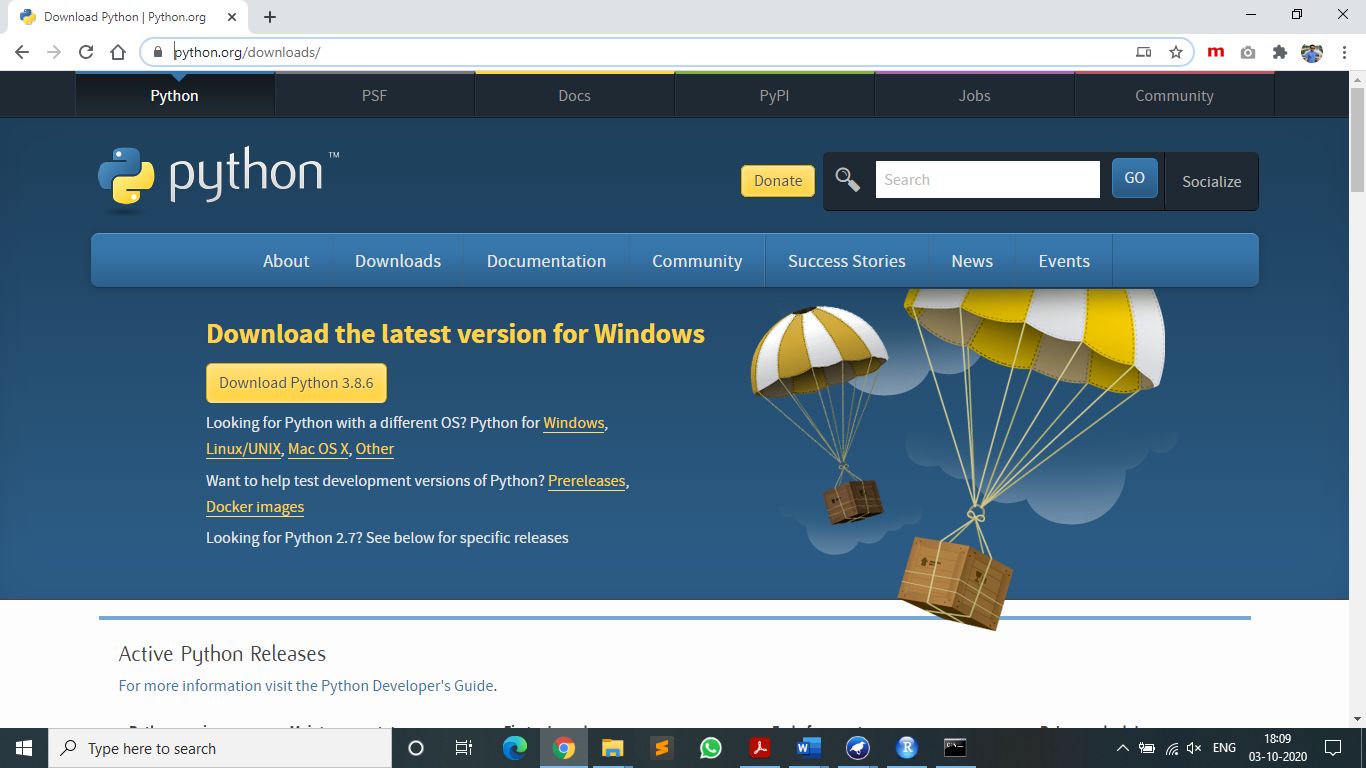
-Object Oriented Approach.

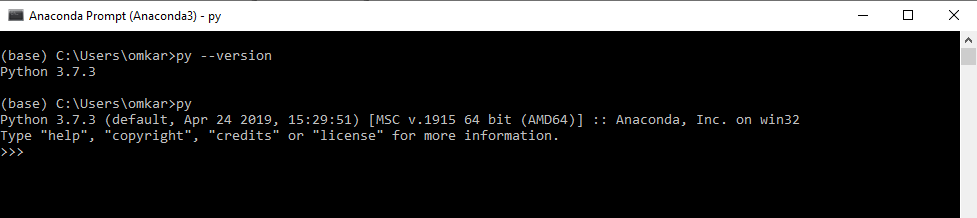
-Highly Portable.

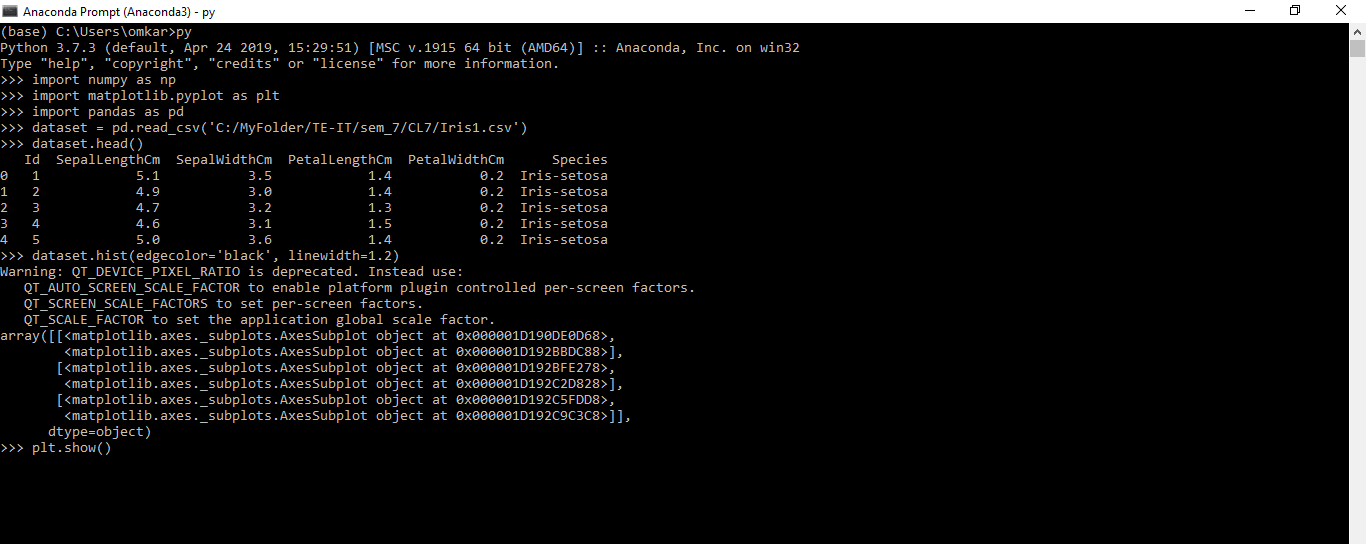
-Highly Dynamic.

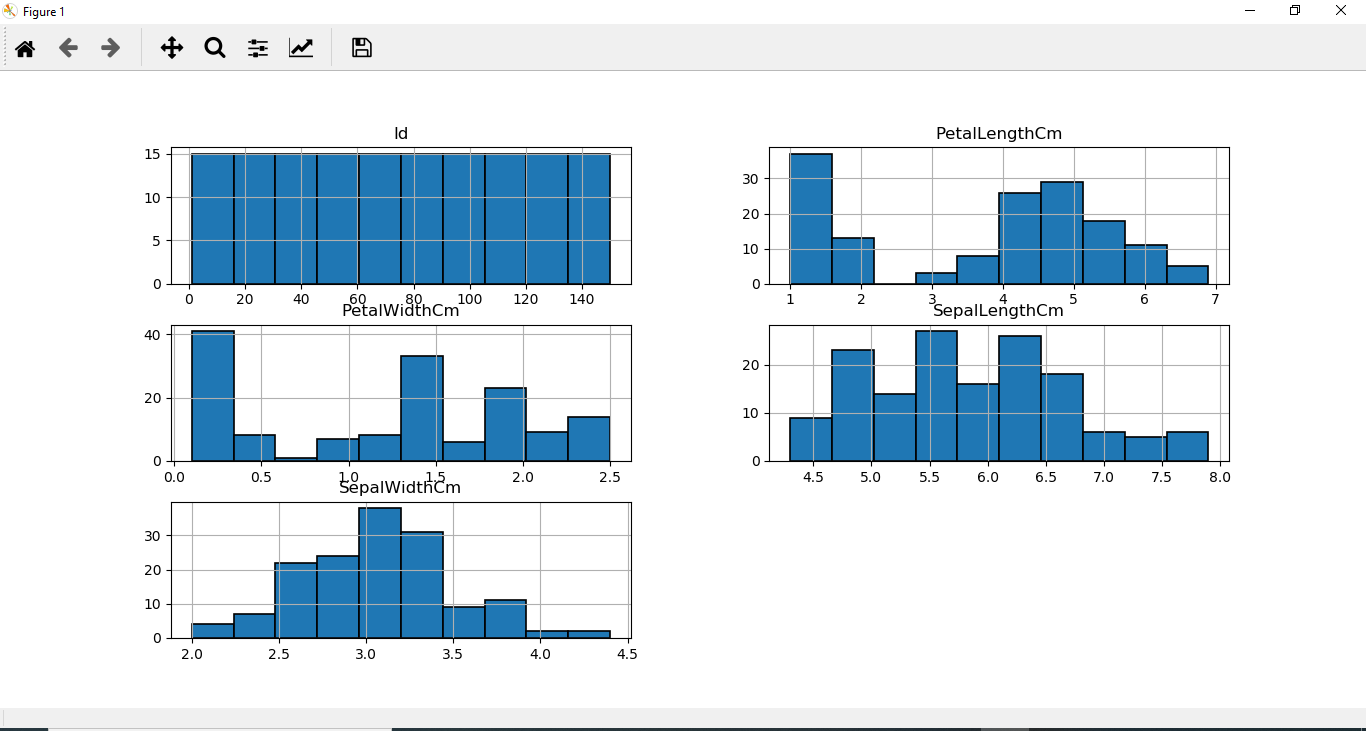
-High -Level Language.

* **Screenshots**

****

****

****



* **Advantages of Python:**

**-** Improved Productivity.

-Interpreted Language.

-Dynamically Typed.

-Free and Open Source.

-Vast Libraries Support.

* **Disadvantages of Python:**

**-** Slow speed

-Not Memory Efficient

-Weak in Mobile Computing

* **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.

**CONCLUSION:**

Downloaded the open source softwares Python, RStudio and WEKA. Studied the distinct features and functionality of these software platforms. Found WEKA easier to learn but there are some limitations in case of Graphical Representations, Modifying the dataset etc. R is difficult to learn for novice but its Graphical Representation is better than WEKA.