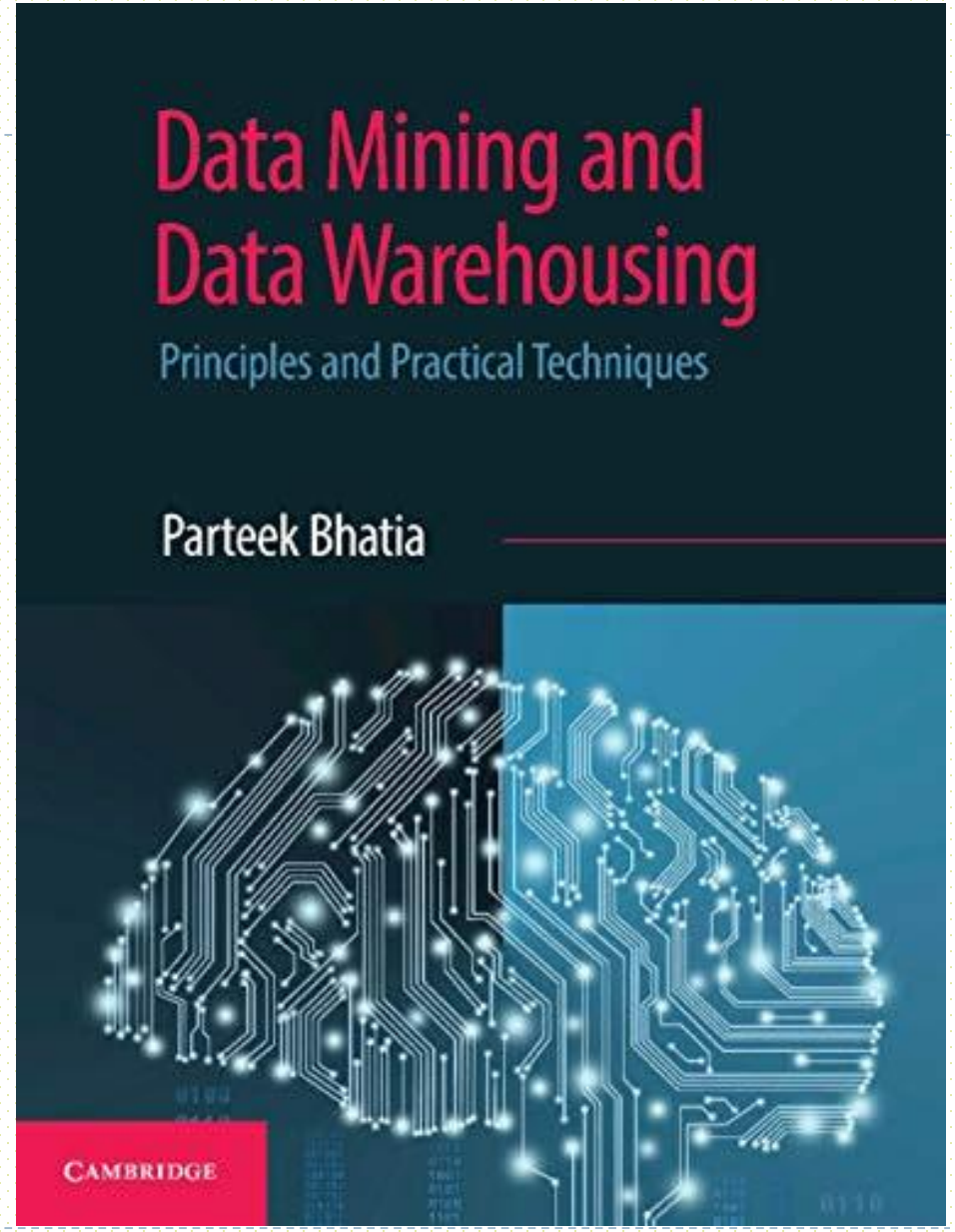
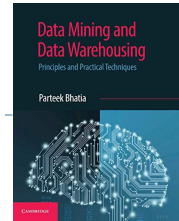
Chapter 3
Beginning with Weka and R language



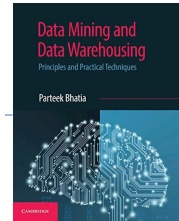
CHAPTER OBJECTIVES

1. To learn to install Weka and the R language
2. To demonstrate the use of Weka software
3. To experiment with Weka on the Iris dataset
4. To introduce basics of R language
5. To experiment with R on the Iris dataset



WEKA

- Weka is an open-source software under the GNU General Public License System. It was developed by the Machine Learning Group, University of Waikato, New Zealand.
- Although named after a flightless New Zealand bird, '***WEKA' stands for Waikato Environment for Knowledge Analysis.***
- The system is written using the object oriented language Java.
- Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization



Installing WEKA

Weka 3 - Data Mining wi

The University of Waikato [NZ] | https://www.cs.waikato.ac.nz/ml/weka/downloading.html

Apps Bookmarks Text Mining - RDataM Documents - RDataM Create Twitter Sentim Made Easy Hand writ Twitter Authenticatio R: Text Mining on Tw Other bookmarks

Downloading and installing Weka

There are two versions of Weka: Weka 3.8 is the latest stable version, and Weka 3.9 is the development version. For the bleeding edge, it is also possible to download nightly snapshots.

Stable versions receive only bug fixes, while the development version receives new features. Weka 3.8 and 3.9 feature a package management system that makes it easy for the Weka community to add new functionality to Weka. The package management system requires an internet connection in order to download and install packages.

Note (1) for users upgrading from Weka 3.7 to Weka 3.8 or later: if the Weka 3.8 package manager does not start up, please delete the file `installedPackageCache.ser` in the `packages` folder that resides in the `wekafiles` folder in your user home.

Note (2) for users upgrading from Weka 3.7 to Weka 3.8 or later: serialized models created in 3.7 are not compatible with 3.8. We have a **model migrator** tool that can migrate some models to be compatible with 3.8.0. One exception is RandomForest, which can be migrated up to 3.7.13 but no further. Usage is as follows:

java -cp <path to modelMigrator.jar>:<path to weka.jar> weka.core.ModelMigrator -i <path to old serialized weka model> -o <upgraded model file name>

If your computer has a display that has a high pixel density, and you are using Windows, Weka's user interfaces may not be scaled appropriately and appear tiny. Installing Java 9 solves this problem. Alternatively, in the Program menu of Weka's GUIChooser, go into Settings, and select WindowsLookAndFeel from the "Look and feel for UI dropdown" menu. This may be necessary because some Weka packages currently do not work as expected with Java 9 (e.g., RPlugin).

- **Snapshots**

  Every night a snapshot of the Subversion repository is taken, compiled and put together in ZIP files. For those who want to have the latest bugfixes, they [illegible] **here**.

- **Stable version**

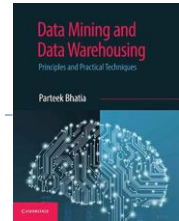
  Weka 3.8 is the latest stable version [illegible] ves bug fixes only, although new features may become available in packages. The[illegible] ng and installing it on your system:

Click here to download

Windows

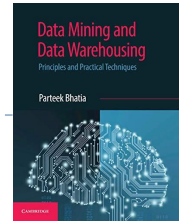
Click **here** to download a self-extracting executable for 64-bit Windows that includes Oracle's 64-bit Java VM 1.8 (weka-3-8-1jre-x64.exe; 112.0 MB)

11:20 AM 11/2/2017

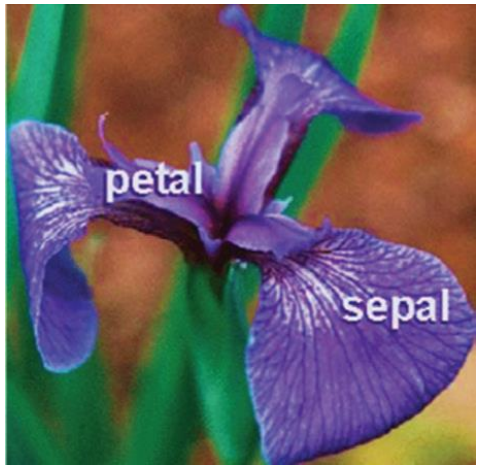


Understanding Fisher's Iris Flower dataset

- Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species. This dataset contains 50 samples of each of the three species, for a total of 150 samples.
- Anderson performed measurements on the three Iris species **(i.e., *Setosa, Versicolor, and Virginica)* using four iris dimensions, namely, Sepal length, Sepal width, Petal length, and Petal width. He had observed that species of the flower could be identified on the basis of these four parameters.**



Understanding Fisher's Iris Flower dataset



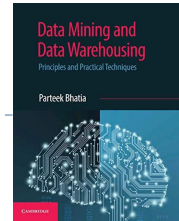
Sample

Class

	Input Attributes				Output Attribute
Instance No.	Sepal Length	Sepal Width	Petal Length	Petal Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5	3.6	1.4	0.2	setosa

Numerical

Nominal



Preparing the Dataset

- The preferred Weka dataset file format is an **Attribute Relation File Format (ARFF) format.**
- An ARFF file is an ASCII text file that describes a list of instances sharing a set of attributes.

Comment

Dataset Name

Attributes

Class Variable

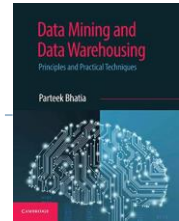
Data values

```
%
% Summary Statistics:
%                Min  Max   Mean    SD   Class Correlation
%     sepal length: 4.3  7.9   5.84  0.83    0.7826
%      sepal width: 2.0  4.4   3.05  0.43   -0.4194
%     petal length: 1.0  6.9   3.76  1.76    0.9490  (high!)
%      petal width: 0.1  2.5   1.20  0.76    0.9565  (high!)
%
% 9. Class Distribution: 33.3% for each of 3 classes.

@RELATION iris

@ATTRIBUTE sepallength  REAL
@ATTRIBUTE sepalwidth   REAL
@ATTRIBUTE petallength  REAL
@ATTRIBUTE petalwidth   REAL
@ATTRIBUTE class        {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
```

Exploring WEKA

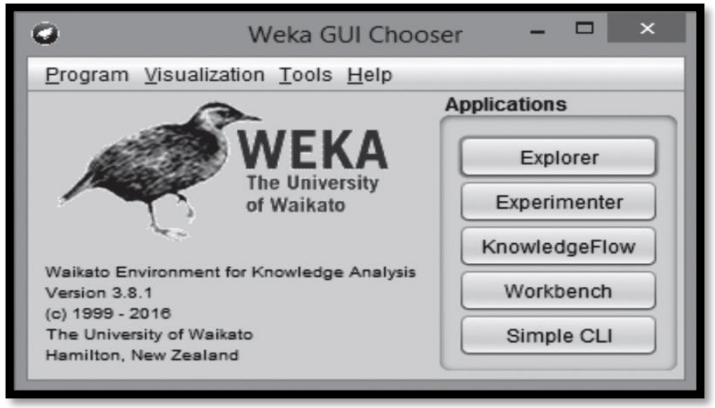
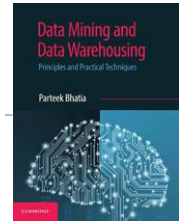
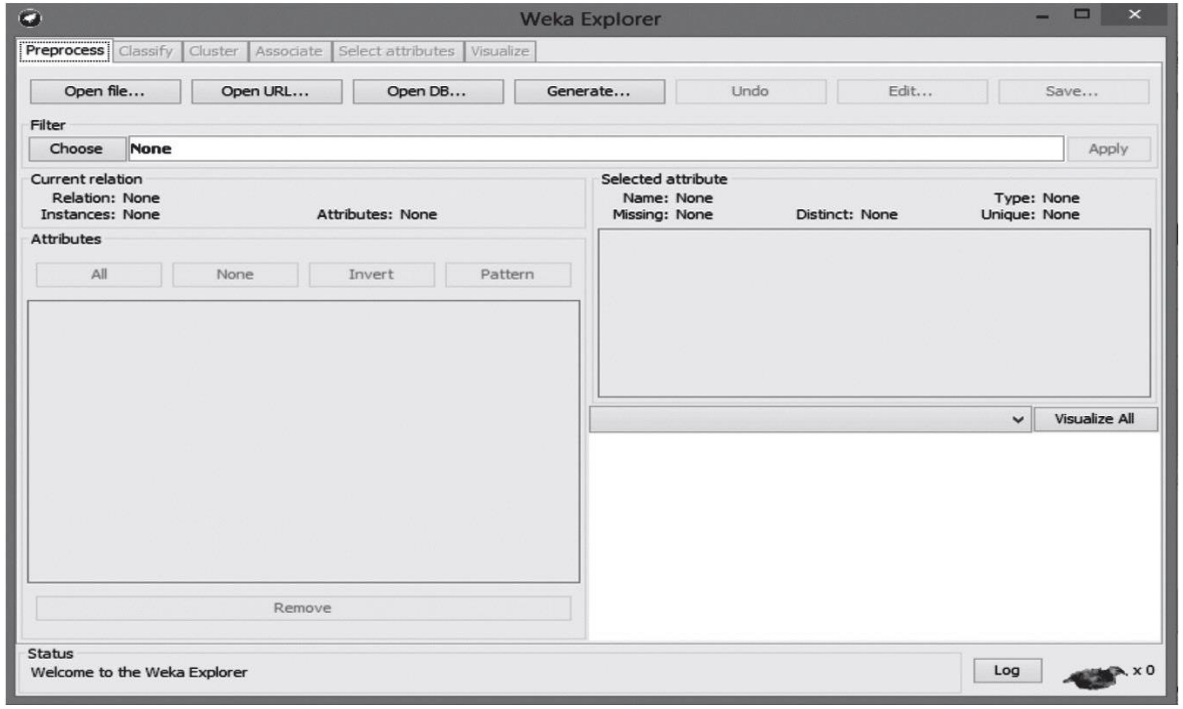


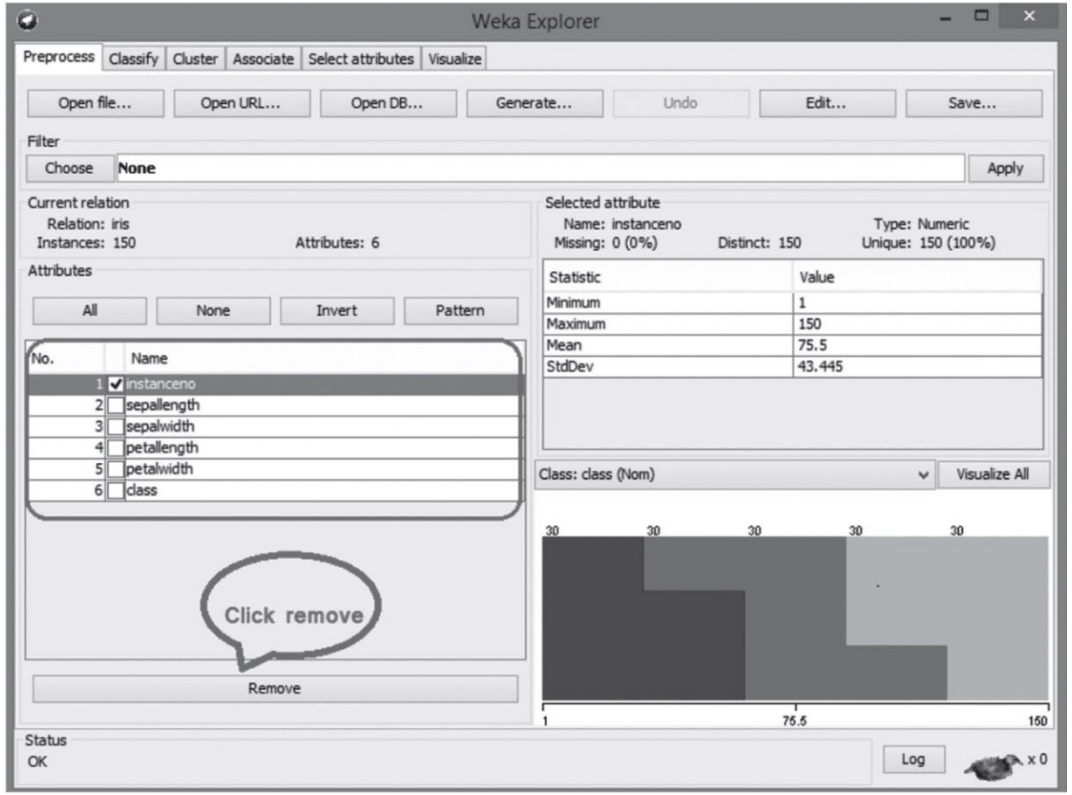
Table 3.1 WEKA GUI applications

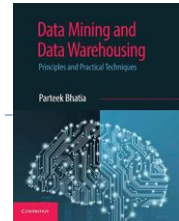
Application	*Description*
Explorer	It is an environment for exploring data.
Experimenter	This interface is for designing experiments with your selection of algorithms and datasets, running experiments and analyzing the results.
Knowledge Flow	It is a Java-Beans based interface to design configurations for streamed data processing.
Workbench	It is a unified graphical user interface that combines the other three such as Explorer, Experimenter and Knowledge Flow (and any plugins that the user has installed) into one application.
Simple CLI	It provides a simple command-line interface and allows direct execution of Weka commands.



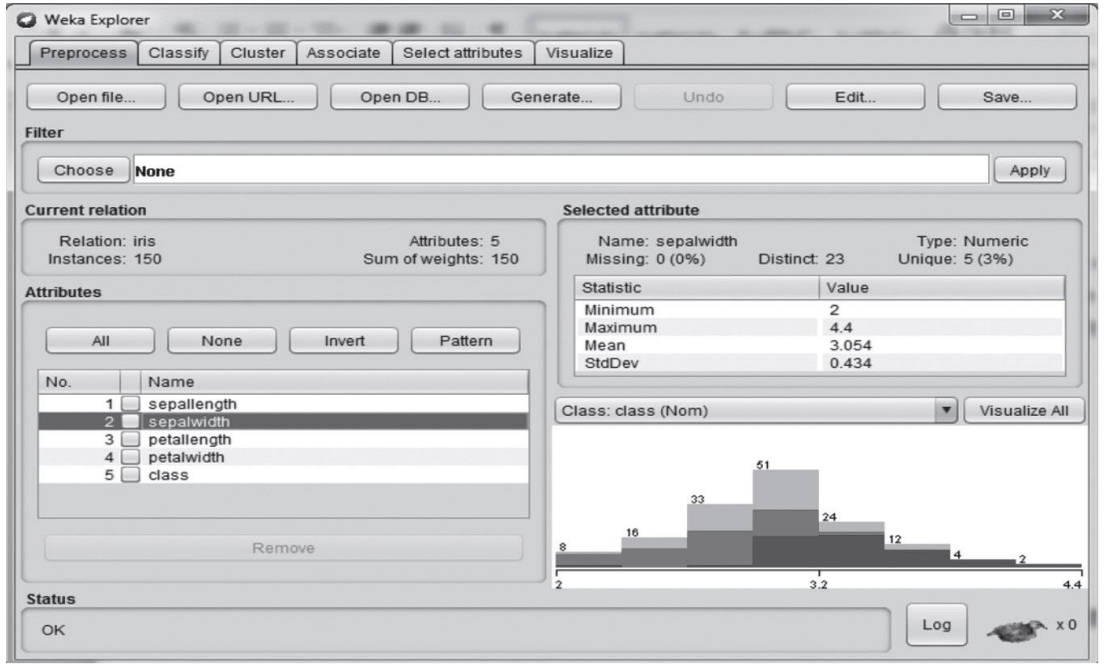
Loading Data

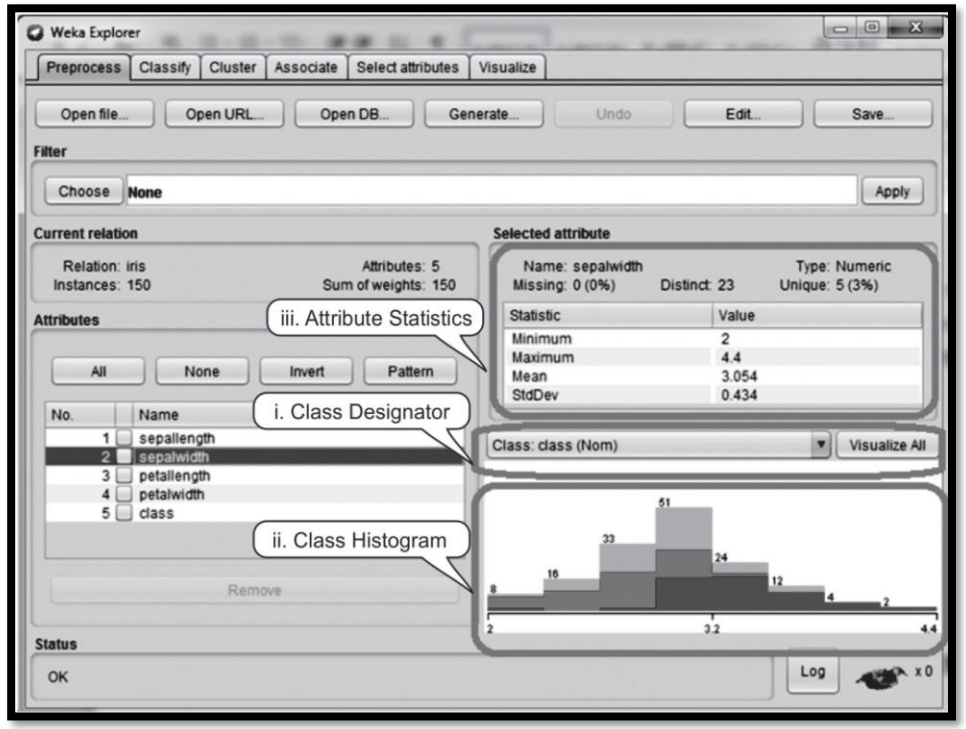


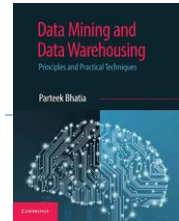




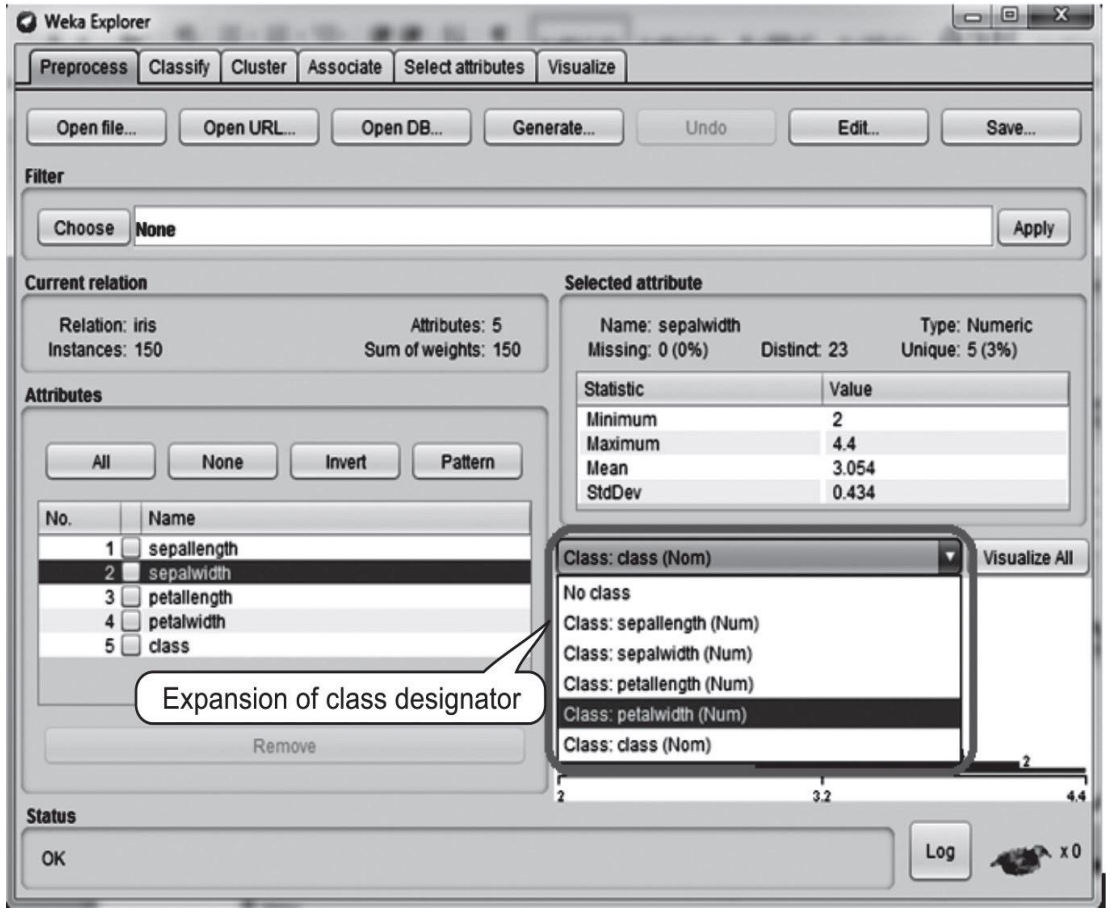
Loading Data

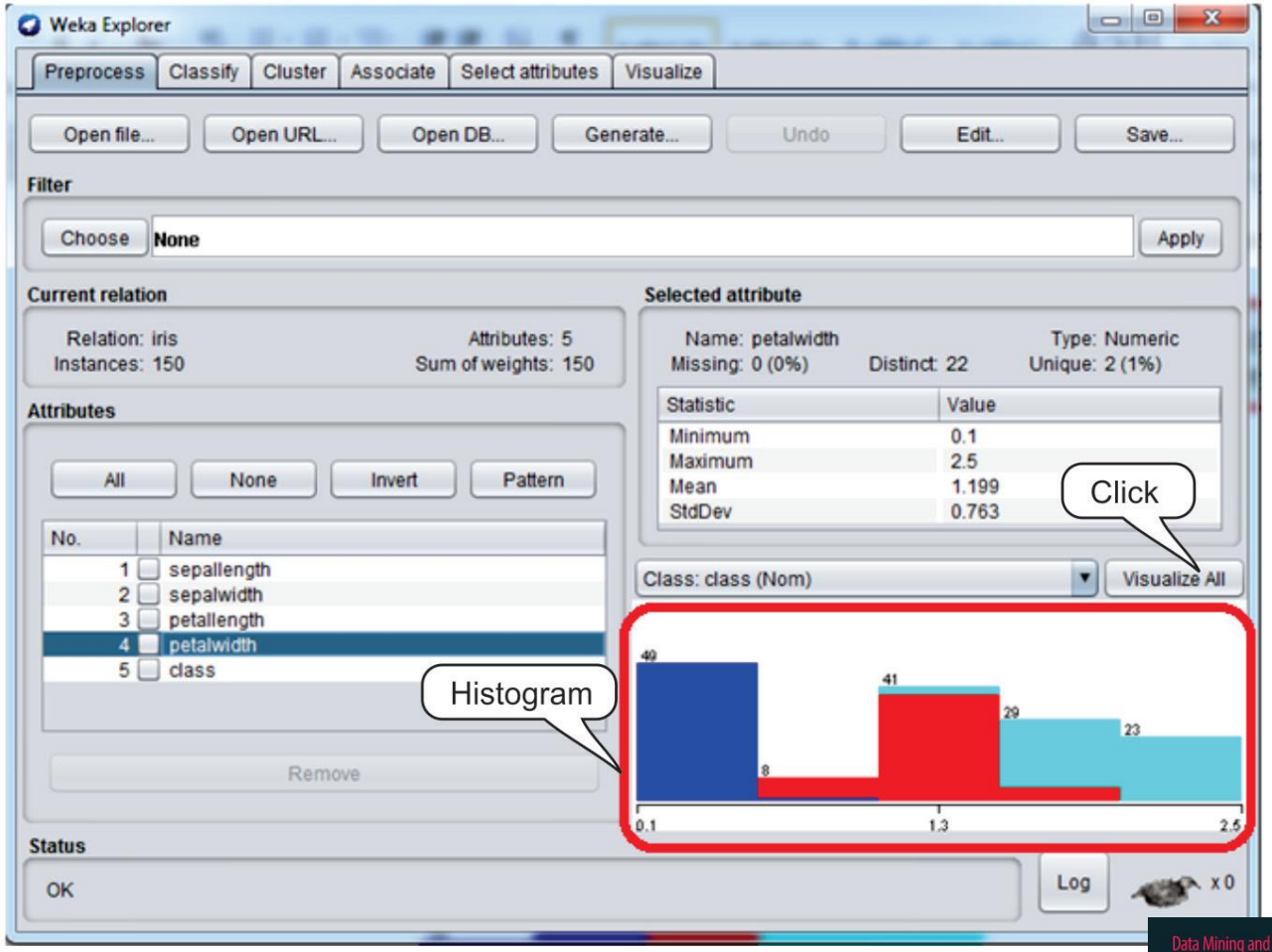


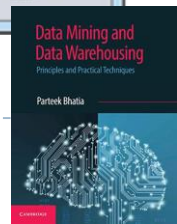




Loading Data

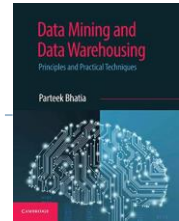






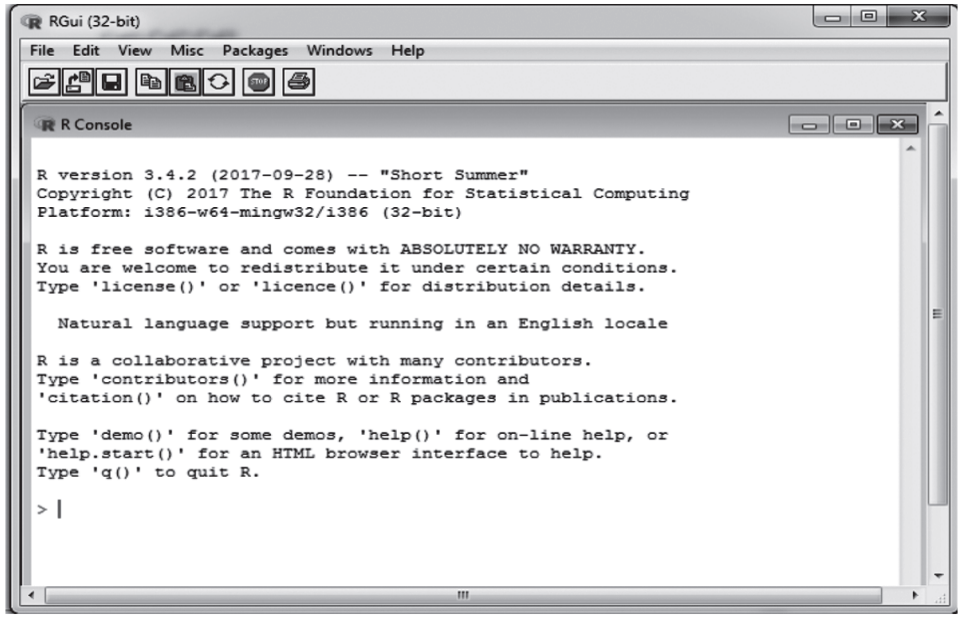
Introduction to R

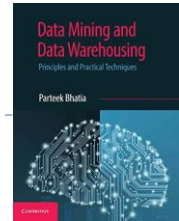
- R is a programming language for statistical computing and graphics.
- It was named R on the basis of the first letter of first name of the two R authors (Robert Gentleman and Ross Ihaka).
- It was developed at the University of Auckland in New Zealand. R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems like Linux, Windows and Mac.



Installing R

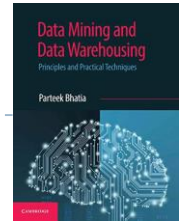
- R can be downloaded from one of the mirror sites available at:
  - ***http://cran.r-project.org/mirrors.html***



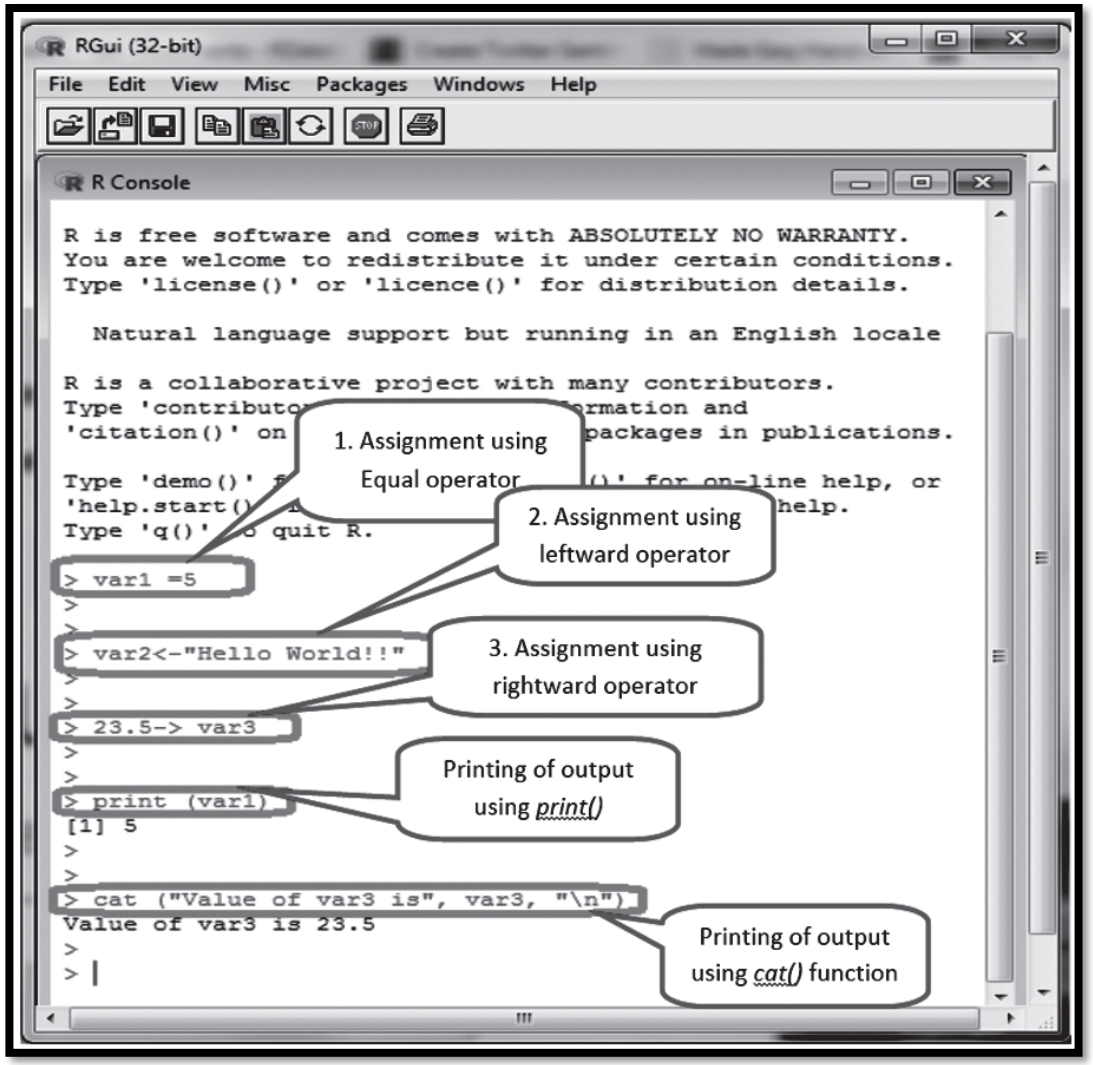


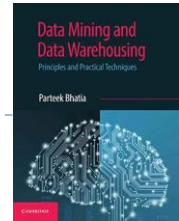
Variable assignment & o/p printing in R

- In R, a variable name consists of letters, numbers and the dot or underline characters. The variable name starts with a letter or the dot not followed by a number. The variables can be assigned values using leftward, rightward and equal to operator. The values of the variables can be printed using *print( ) or cat( ) function. cat( ) function combines multiple items into a continuous print output.*



Example

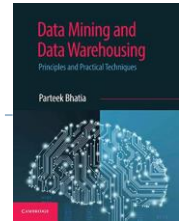

RGui (32-bit)
File Edit View Misc Packages Windows Help
R Console
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
> var1 =5
> var2<-"Hello World!!"
> 23.5-> var3
> print (var1)
[1] 5
> cat ("Value of var3 is", var3, "\n")
Value of var3 is 23.5
1. Assignment using Equal operator
2. Assignment using leftward operator
3. Assignment using rightward operator
Printing of output using print()
Printing of output using cat() function


Data Mining and Data Warehousing
Parteek Bhatia

Data Types in R

Table 3.2 Description about basic data types

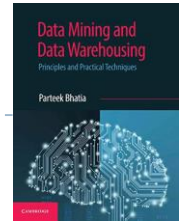
Data Type	*Description*	*Examples*
Character	A character object is used to represent string values.	'A', 'I am learning programming'
Numeric	Numeric stores the real or decimal values.	10, 25.2
Integer	Integer is used to store integer values.	2L (the L tells R to store this as an integer)
Logical	A logical value is created via comparison between variables.	TRUE, FALSE
Complex	A complex value in R is defined via the pure imaginary value *i*.	2+5*i* (complex numbers with real and imaginary parts)



Basic Operators in R

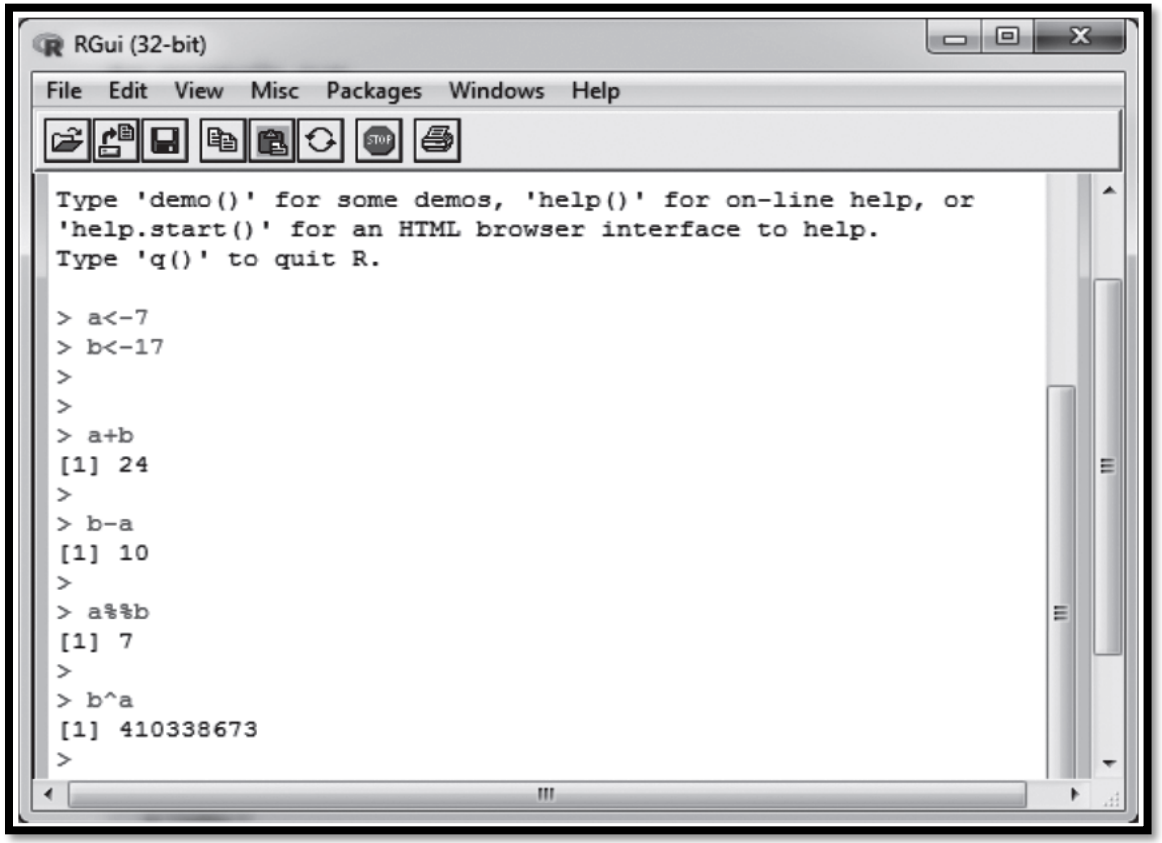
Table 3.3 Summary about basic operators of R

Type	*Operators*
Arithmetic	+, -, *, %%, ^
Relational	<. >, <=, >=, !=
Logical	&, |, &&, ||, !
Assignment	=, <-, ->

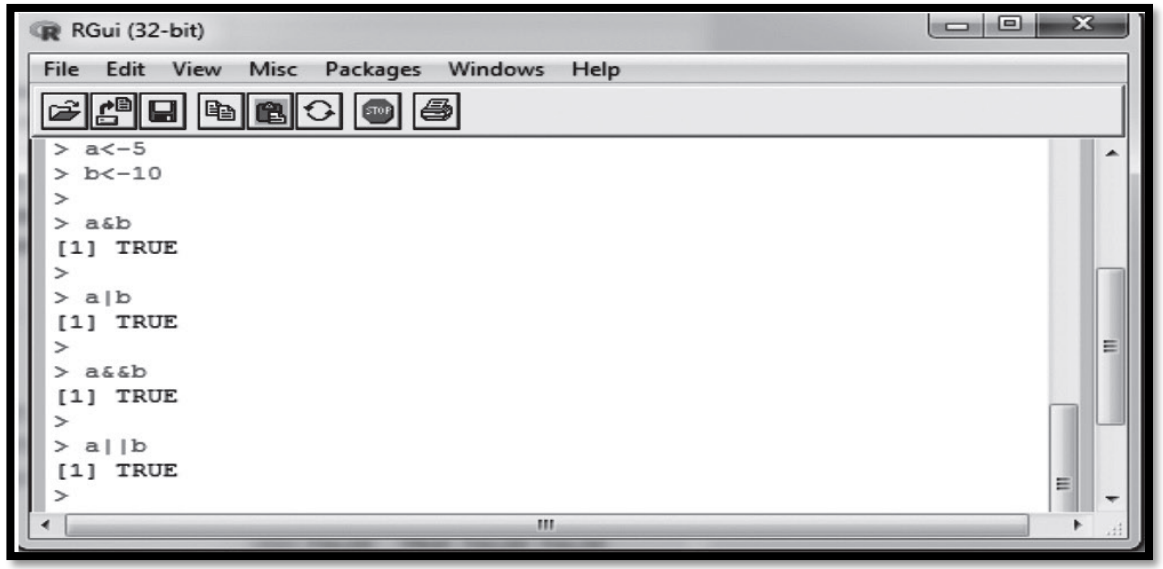


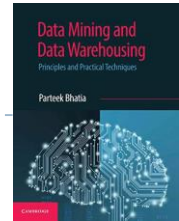
Operators in R

Arithmetic operators



Logical operators

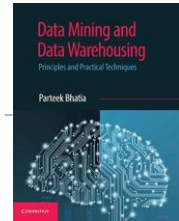




Machine Learning Packages in R

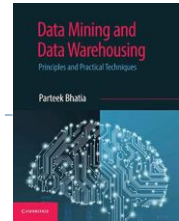
Table 3.4 Some of the important machine learning packages

Sr. No.	Package Name	Description
1.	e1071	This package is used for implementing Naïve Bayes (conditional probability), SVM, Fourier Transforms, Bagged Clustering, Fuzzy Clustering, etc.
2.	CORElearn	It is used for classification, regression, feature evaluation and ordinal evaluation.
3.	randomForest	It is used to create large number of decision trees and then each observation is inputted into the decision tree.
4.	Arules	This package is used for Mining Association Rules and Frequent Itemsets.
5.	MICE	This package is used to assign missing values by using multiple techniques, depending on the kind of data.
6.	RPART (Recursive Partitioning and Regression Trees)	It is used to build classification or regression models using a two stage procedure and the resultant models are represented in the form of binary trees.
7.	nnet	This package is used for Feed-forward Neural Networks and Multinomial Log-Linear Models.



Loading of Data in R

```
>library(gdata)        # load gdata package
>mydata = read.xls("mydata.xls") # read from first sheet
Or
>mydata = read.csv("mydata.csv") # read from csv format
Or
>mydata = read.arff("mydata.arff") # read from arff format
```

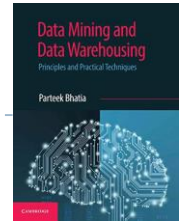



Working with the iris dataset in R

```
>library (datasets)          # load datasets package
> data (iris)                # load dataset
> names (iris)               # display attribute names
```

```
> summary (iris)
> summary (iris$Sepal.Width)
```

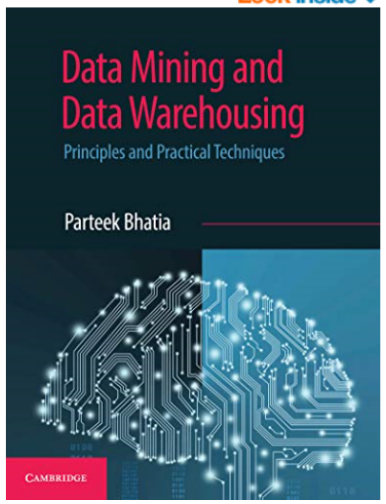
```
> View(iris) #To view the dataset instances
```



Reference

https://www.amazon.in/Data-Mining-Warehousing-Principles-Techniques-ebook/dp/B07R8VPLSH

Look inside



Data Mining and Data Warehousing: Principles and Practical Techniques [Print Replica] Kindle Edition

by Parteek Bhatia (Author)

11 ratings

See all 2 formats and editions

Kindle Edition	Paperback
₹ 568.10	₹ 598.00 ✓prime

Read with Our Free App

37 New from ₹ 499.00

Written in lucid language, this valuable textbook brings together fundamental concepts of data mining and data warehousing in a single volume. Important topics including information theory, decision tree, Naïve Bayes classifier, distance metrics, partitioning clustering, associate mining, data marts and operational data store are discussed comprehensively. The textbook is written to cater to the needs of undergraduate students of computer science, engineering and information

Read more

Length: 506 pages Language: English Format: Print Replica

Available on these devices

Due to its large file size, this book may take longer to download

Kindle Price: ₹ 568.10
inclusive of all taxes

includes free wireless delivery via ***Amazon Whispernet***

Sold by: Amazon Asia-Pacific Holdings Private Limited

Add to cart

Buy now with 1-Click®

Using Amazon Pay balance. Why?

Deliver to:

Parteek's 10th Android Device

Enter a promotion code or Gift Card

Send a free sample

Deliver to:

Parteek's 10th Android Device

Add to Wish List

Share <Embed>

Follow the Author

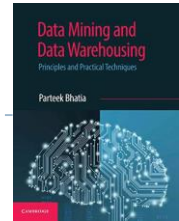
Parteek Bhatia ✓ Following

Customers who bought this item also bought

Page 1 of 3

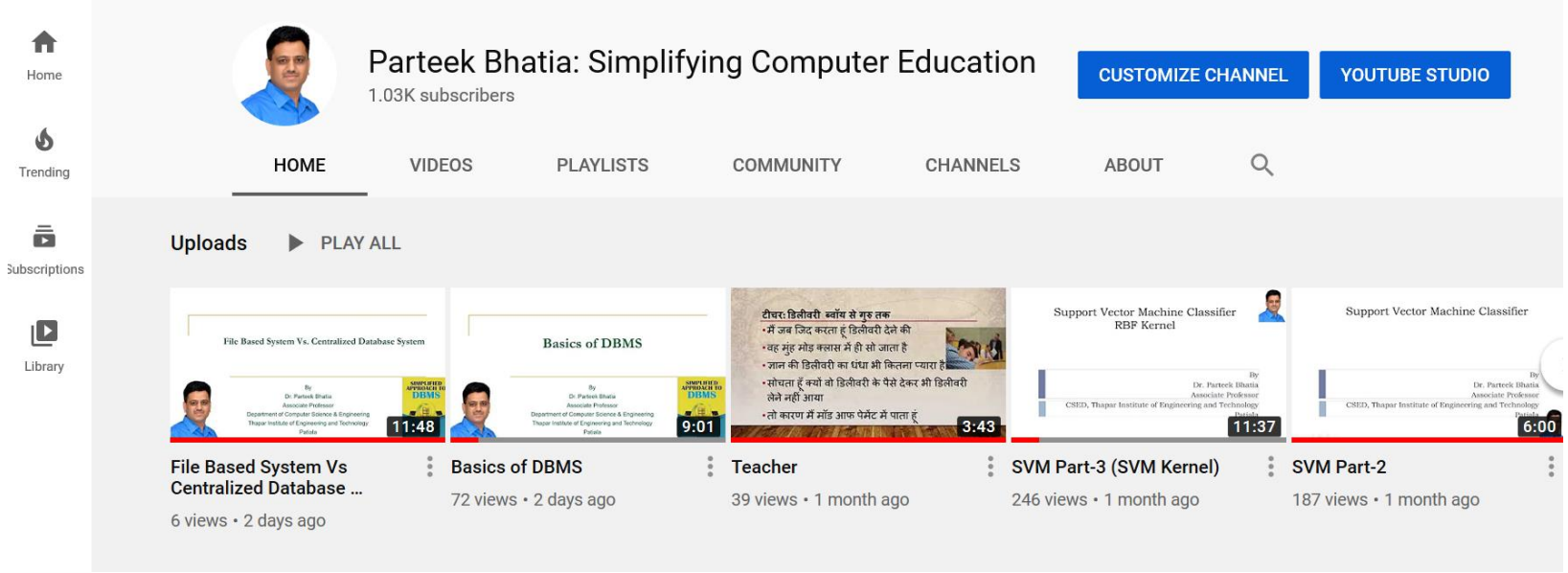
Type here to search

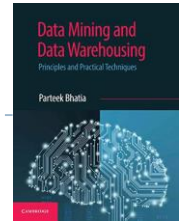
ENG 21:56 12-01-2020



For more information

- Subscribe to YouTube Channel from the Author
  - To receive latest video tutorials on Data Mining, Machine Learning, DBMS, Big Data, NoSQL and many more.
- https://www.youtube.com/user/parteekbhatia





Free Online on SQL at Udemy

Udemy

ROYAL ACADEMY OF ENGINEERING

Funded

THAPAR INSTITUTE OF ENGINEERING & TECHNOLOGY (Deemed to be University)

Simplified Approach to SQL

QUICK START
Ace Interviews and College Exams in 1 month

BI - LINGUAL
Explanation of each of the topic in English and Hindi

INTERACTIVE
Discussion forums and direct messaging to instructor

RELAXING
Study the course at your own pace

STUDY ANYWHERE
Study on your smartphone with udemy app

COURSE CONTENT

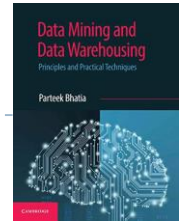
Week 1 : *Introduction to SQL and performing basic operations with SQL.*
Week 2 : *Creation of Tables with Integrity constraints.*
Week 3 : *Table Alterations and Joins.*
Week 4 : *Grouping of Data*

ABOUT THE INSTRUCTOR

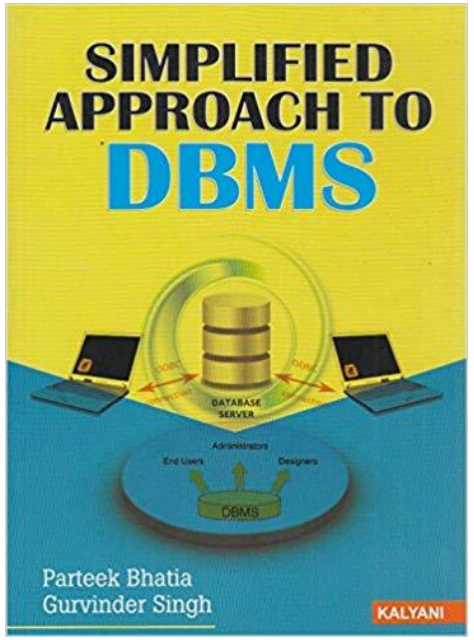
Dr. Parteek Bhatia is Associate Professor in the Department of Computer Science and Engineering at Thapar Institute of Engineering and Technology, Patiala. He has more than 18 years of academic experience.He has authored several books in various areas of computer science. His book - Simplified approach to DBMS is one of the bestseller. Currently, he is working on plethora of Projects which are funded by Department of Science and Technology, CSIR and other funding agencies of India.

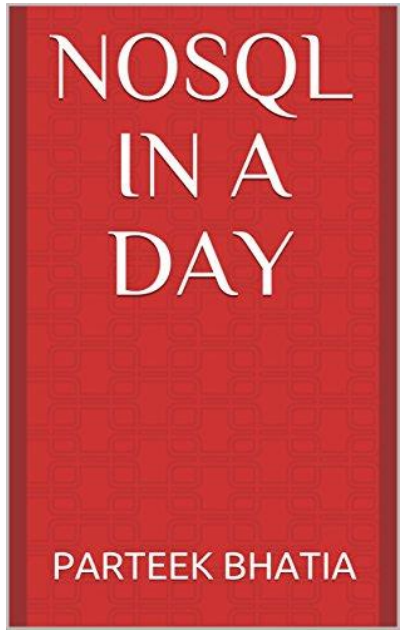
QR CODE

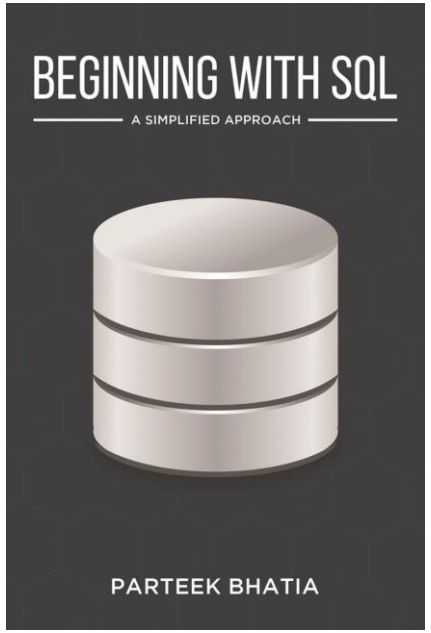
COURSE LINK: https://www.udemy.com/learn-sql-in-a-simplified-manner/

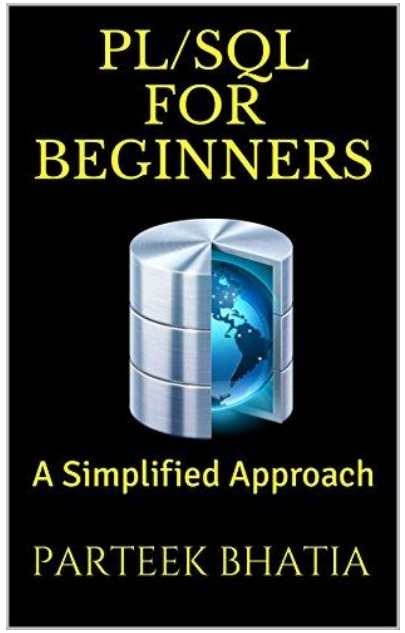


Books from the Same Author









For more information visit: www.parteekbhatia.com

ABOUT THE AUTHOR

Dr. Bhatia is an Associate Professor in the Department of Computer Science and Engineering at Thapar Institute of Engineering and Technology, Patiala. He has more than twenty years of teaching experience and has published papers in journals. His current research includes natural language processing, machine learning and human-computer interface. He has taught courses including data mining and data warehousing, big data analysis and database management system at undergraduate and graduate levels. He also runs online courses on the Udemy portal.

www.parteekbhatia.com

For book related queries : parteek.bhatia@gmail.com