EX.NO:1 STATISTICAL ANALYSIS USING R

AIM:

To perform statistical analysis using R language.

PROCEDURE:

Introduction to statistical analysis with R:

Statistical Analysis with R is one of the best practices which the statistician, data analysts, and data scientists do while analyzing statistical data. R language is a popular open-source programming language that extensively supports built-in packages and external packages for statistical analysis. R language natively supports basic statistical calculations for exploratory data, and advanced statistics for predictive data analysis Statistical analysis with R is an important part of identifying data patterns based upon the statistical rules and business constraints. Due to the simplicity of R syntax and flexibility of using advanced packages. R language is preferred for Statistical Analysis.

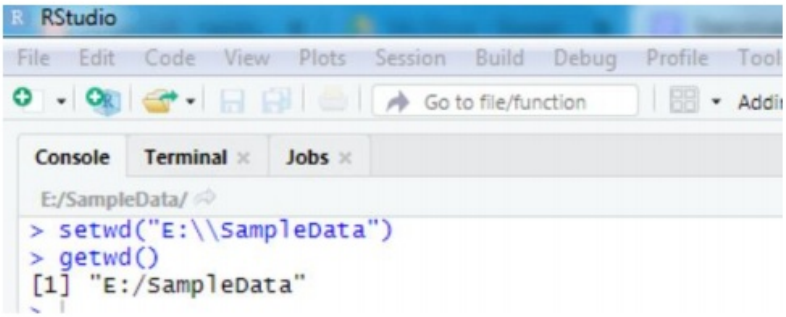
How to perform statistical analysis with R language:

- To start with statistical data analysis with R, the business requirement needs to be clear to find the data patterns from the available data.
- The R language needs to be installed on the system
    - R can be installed in Windows, Linux, and MAC OS X.
    - The installable file for R can be downloaded from **https://cran.r-project.org/**.
- Next, the IDE such as R Studio needs to be installed on the system.
    - R Studio provides GUI support along with some enterprise-ready features like Syntax hiliting, debugging, packages, and workspace management.

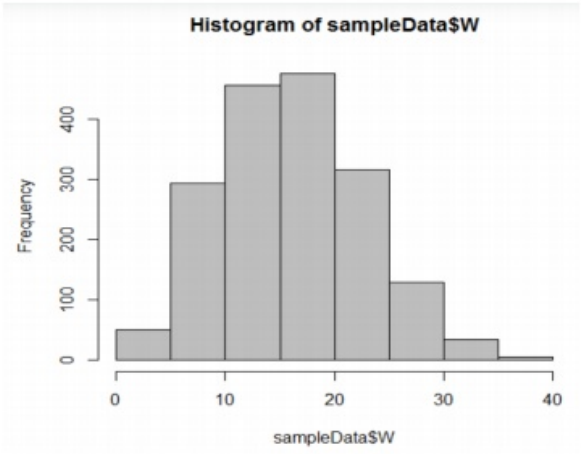
- Once the Environment is ready, the next step is to import the data set to R workspace.
    - For Example, we will import a .csv file to R studio for Statistical analysis.
- R Studio can be downloaded and installed from **https://posit.co/**
    - Once the R studio is installed, it can be directly used to develop R script which will work on the installed version of the R language.
    - We will be downloading an open-source data set from **https://www.kaggle.com/** for this demonstration.
    - The data file we will use is 'cbb.csv' which is college basketball dataset,

The practical approach of statistical analysis with R

- This section will do hands-on using R studio for college basketball dataset.
    - The first step is to set the working directory which will be used as the preferred location to read and write datasets.
    - setwd() is used in R to set the working directory
    - getwd() to check the present working directory
    - Following is a screenshot of R Studio with setwd() and getwd() functions.



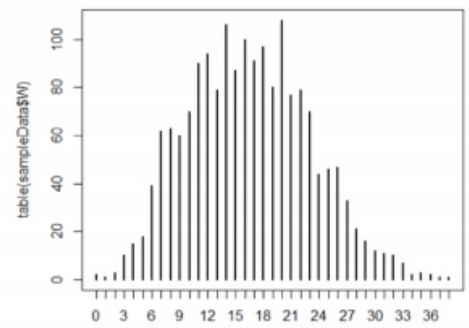
- Next will import the data set using read.csv() command and assign to a data frame called SampleData as the following the syntax.
- Sample data = read.csv("cbb.csv").
- To check the dataset imported correctly and review the few top lines of data use head() command in R



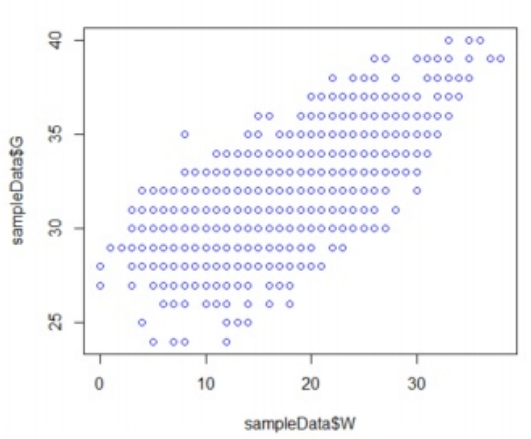
- We can use Table function to create a frequency table which shows the number of frequency of the data in the variable using table(sampleData$W).
- The frequency table shows the value 20 has a maximum frequency in the data. This function is very useful while doing statistical categorical variables.
- Also, we n plat this frequency table using plot function in R using >

```
> table(sampleData$W)

  0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22
  2   1   3  10  15  18  39  62  63  60  70  90  94  79 106  87 100  91  97  80 108  77  79
 23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38
 70  44  46  47  33  21  16  12  11  10   7   2   3   2   1   1
```



- This statistical analysis is a comparison between two variables present in that data set.
- It helps to identify the correlation and patterns between the two variables.
- Symbol '~' is used for bivariate analysis in R.
- This scatter plot represents the graph for bivariate analysis



  - Apart from the Scatter plot, there are several other functions and plots like histograms, line plots, and boxplots are being used for Bivariate data analysis.
  -
- Next, we will discuss the t-test which is the statistical hypothesis testing process using R.
  - t,test() function used in R to process the t-test
  - We will use G variable data of data frame sample data for t-test
  - test(sampleDat$G) is the syntax we will apply on the R Studio console.
  - T-test shows the statistical inferences and the confidence interval .as outcomes.
  - The p-value is the probability value significant to the null hypothesis. And the percentage value is the confidence interval.

Importance of Statistical Analysis with R language:

- R is a reliable programming language for Statistical Analysis.
- It has a wide range of statistical library support like T-test, linear regression, logistic regression, time-series data analysis.

- It is a scripting language, which helps statisticians and data scientists to develop code and test individual statistical models for efficient data analysis.
- The code written in R for statistical analysis is easy for interpretation and sharable to other stack holders of the organization and coworkers.
- Being a popular and well-structured Language, R has several code reusable components and libraries available to get started with statistical analysis of an input dataset.
- R language includes various build-in datasets for learning and creating a proof of concept before using actual business data for statistical analysis.
- R comes with very good data visualization features supporting potting and graphs using graphical packages like ggplot2.

RESULT:

Thus the statistical analysis of R language has been performed.

EX.NO:2A **STUDY ON WEKA TOOL**

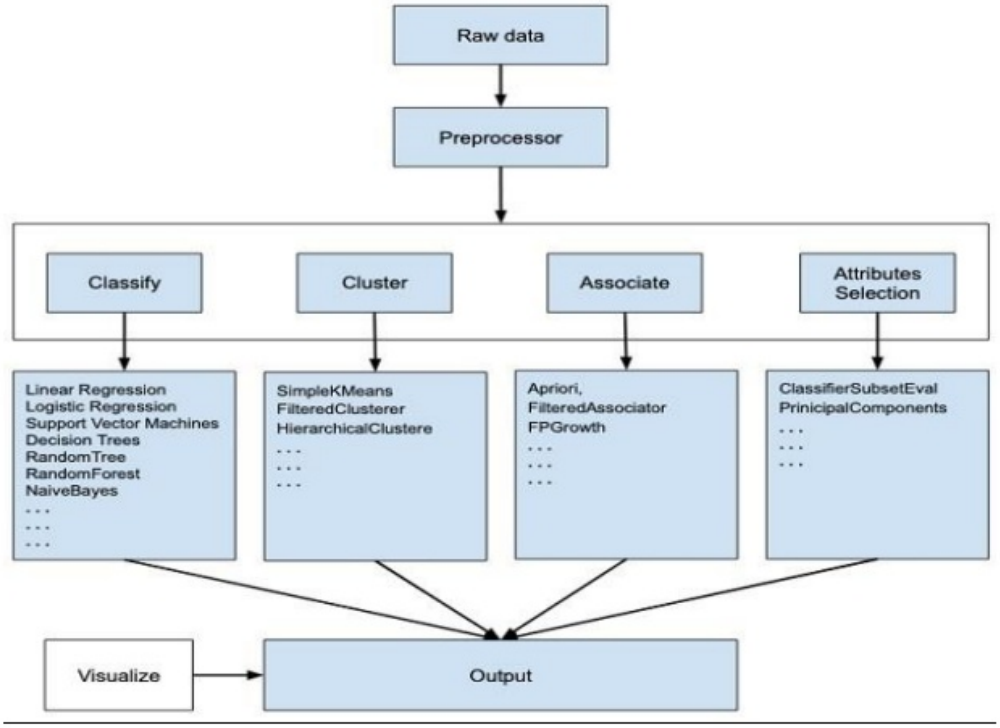
AIM:

To study on WEKA tool.

PROCEDURE:

WEKA:

WEKA - an open source software provides tools for data preprocessing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problems.



First, you will start with the raw data collected from the field. This data may contain several null values and irrelevant fields. You use the data preprocessing tools provided in WEKA to cleanse the data.

Then, you would save the preprocessed data in your local storage for applying ML algorithms.

Next, depending on the kind of ML model that you are trying to develop you would select one of the options such as **Classify, Cluster**, or **Associate**. The **Attributes Selection** allows the automatic selection of features to create a reduced dataset.

Note that under each category, WEKA provides the implementation of several algorithms. You would select an algorithm of your choice, set the desired parameters and run it on the dataset.

Then, WEKA would give you the statistical output of the model processing. It provides you a visualization tool to inspect the data.

The various models can be applied on the same dataset. You can then compare the outputs of different models and select the best that meets your purpose.

Thus, the use of WEKA results in a quicker development of machine learning models on the whole.

Weka supports several standard data mining tasks, specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. Input to Weka is expected to be formatted according to the Attribute-Relational File Format and filename with the .arff extension.

All Weka's techniques are predicated on the assumption that the data is available as one flat file or relation, where a fixed number of attributes describes each data point (numeric or nominal attributes, but also supports some other attribute types). Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. Weka provides access to deep learning with Deeplearning4j.

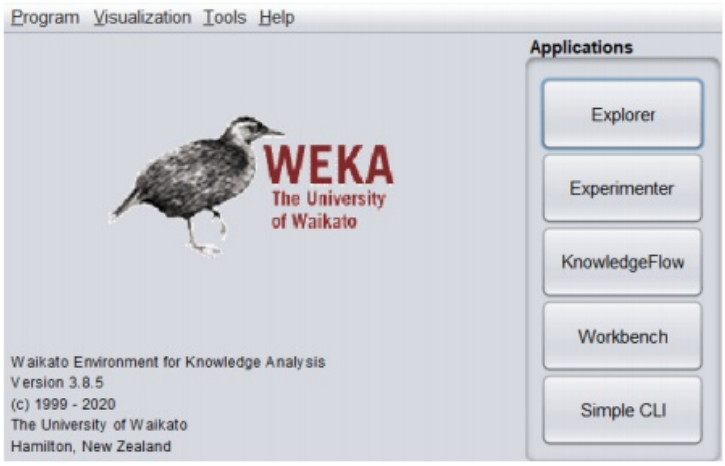
It is not capable of multi-relational data mining. Still, there is separate software for converting a collection of linked database tables into a single table suitable for processing using Weka. Another important area currently not covered by the algorithms included the Weka distribution in sequence modelling.

Features of WEKA:

- Select attributes
- Visualize
- Preprocess
- Classify
- Cluster
- Associate

Requirements and Installation of WEKA:

We can install WEKA on Windows, MAC OS, and Linux. The minimum requirement is Java 8 above for the latest stable versions of Weka.



- The Explorer is the central panel where most data mining tasks are performed. We will further explore this panel in upcoming sections.
- The tool provides an Experimenter In this panel, we can run experiments and also design them.
- WEKA provides the KnowledgeFlow panel. It provides an interface to drag and drop components, connect them to form a knowledge flow and analyze the data and results.
- The Simple CLI panel provides the command line powers to run WEKA. For example, to fire up the ZeroR classifier on the arff data, we'll run from the command line:

RESULT:

Thus the study on WEKA tool and its component has been completed successfully.