ITA 0451 - STATISTICS WITH R PROGRAMMING FOR DEEP LEARNING

SAVEETHA SCHOOL OF ENGINEERING

SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES

COURSE CODE / SUBJECT: ITA 0451 / STATISTICS WITH R PROGRAMMING FOR DEEP LEARNING

DAY 1 – LAB ASSESSMENT

Reg No:

Name:

1. Write a R program to take input from the user (name and age) and display the values. Also print the version of R installation.

2. Write a R program to get the details of the objects in memory.

3. Write a R program to create a sequence of numbers from 20 to 50 and find the mean of numbers from 20 to 60 and sum of numbers from 51 to 91.

4. Write a R program to create a vector which contains 10 random integer values between -50 and +50.

5. Write a R program to get the first 10 Fibonacci numbers.

6. Write a R program to get all prime numbers up to a given number (based on the sieve of Eratosthenes).

7. Write a R program to print the numbers from 1 to 100 and print "Fizz" for multiples of 3, print "Buzz" for multiples of 5, and print "FizzBuzz" for multiples of both.

8. Write a R program to extract first 10 english letter in lower case and last 10 letters in upper case and extract letters between 22nd to 24th letters in upper case.

9. Write a R program to find the factors of a given number.

10. Write a R program to find the maximum and the minimum value of a given vector.

11. Write a R program to get the unique elements of a given string and unique numbers of vector.

12. Write a R program to create three vectors a,b,c with 3 integers. Combine the three vectors to become a 3×3 matrix where each column represents a vector. Print the content of the matrix.

13. Write a R program to create a list of random numbers in normal distribution and count occurrences of each value.

14. Write a R program to create three vectors numeric data, character data and logical data. Display the content of the vectors and their type.

15. Write a R program to create a 5 x 4 matrix , 3 x 3 matrix with labels and fill the matrix by rows and 2 × 2 matrix with labels and fill the matrix by columns.

16. Write a R program to create an array, passing in a vector of values and a vector of dimensions. Also provide names for each dimension.

17. Write a R program to create an array with three columns, three rows, and two "tables", taking two vectors as input to the array. Print the array.

18. Write a R program to create a list of elements using vectors, matrices and a function. Print the content of the list.

DAY 2 – LAB ASSESSMENT

Reg No:
Name:
1.Write a R program to create an array of two 3x3 matrices each with 3 rows and 3 columns from two given two vectors. Print the second row of the second matrix of the array and the element in the 3rd row and 3rd column of the 1st matrix.
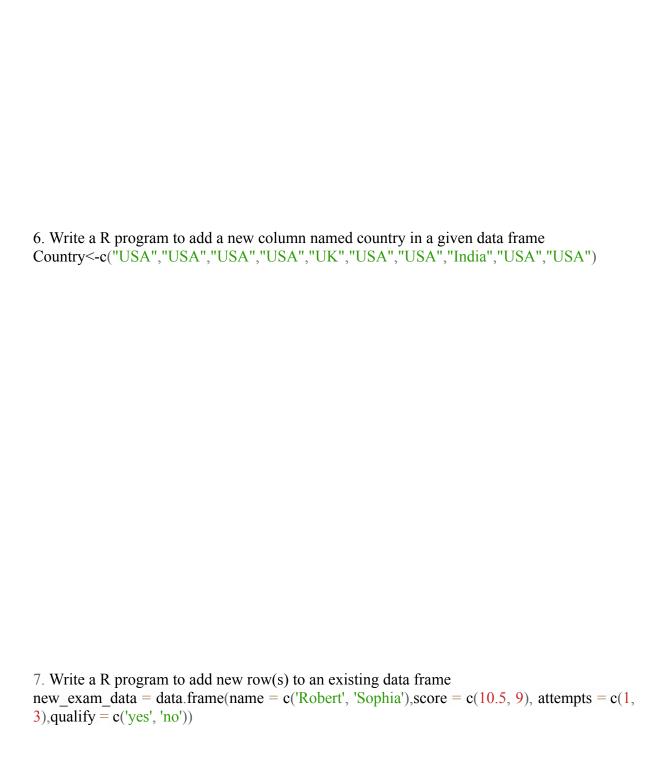
2. Write a R program to combine three arrays so that the first row of the first array is followed by the first row of the second array and then first row of the third array.

3. Write a R program to create an array using four given columns, three given rows, and two given tables and display the content of the array.

4. Write a R program to create a two-dimensional 5x3 array of sequence of even integers greater than 50.

Use Below Data frame from question 5 to 9
exam_data = data.frame(
name = c('Anastasia', 'Dima', 'Katherine', 'James', 'Emily', 'Michael', 'Matthew', 'Laura', 'Kevin', 'Jonas'),
score = c(12.5, 9, 16.5, 12, 9, 20, 14.5, 13.5, 8, 19),
attempts = c(1, 3, 2, 3, 2, 3, 1, 1, 2, 1),
qualify = c('yes', 'no', 'yes', 'no', 'no', 'yes', 'yes', 'no', 'no', 'yes')
)
5. Write a R program to extract 3rd and 5th rows with 1st and 3rd columns from a given data frame

6. Write a R program to add a new column named country in a given data frame
Country<-c("USA","USA","USA","USA","UK","USA","USA","India","USA","USA")

7. Write a R program to add new row(s) to an existing data frame
new_exam_data = data.frame(name = c('Robert', 'Sophia'),score = c(10.5, 9), attempts = c(1, 3),qualify = c('yes', 'no'))

8. Write a R program to sort a given data frame by name and score

9. Write a R program to save the information of a data frame in a file and display the information of the file.

10. Write a R program to call the (built-in) dataset airquality. Check whether it is a data frame or not? Order the entire data frame by the first and second column. remove the variables 'Solar.R' and 'Wind' and display the data frame.

11. Write a R program to create a factor corresponding to height of women data set , which inbuild in R, contains height and weights for a sample of women.

12. Write a R program to extract the five of the levels of factor created from a random sample from the LETTERS (Part of the base R distribution.)

13. **Iris** dataset is a very famous dataset in almost all data mining, machine learning courses, and it has been an R build-in dataset. The dataset consists of 50 samples from each of three species of Iris flowers (Iris setosa, Iris virginica and Iris versicolor). Four features(variables) were measured from each sample, they are the **length** and the **width** of sepal and petal, in centimetres. Perform the following EDA steps .

(i)Find dimension, Structure, Summary statistics, Standard Deviation of all features.

(ii)Find mean and standard deviation of features groped by three species of Iris flowers (Iris setosa, Iris virginica and Iris versicolor)

(iii)Find quantile value of sepal width and length

(iV)create new data frame named iris1 which have a new column name **Sepal.Length.Cate** that categorizes "Sepal.Length" by quantile

(V) Average value of numerical varialbes by two categorical variables: Species and Sepal.Length.Cate:

(vi) Average mean value of numerical varialbes by Species and Sepal.Length.Cate

(vii)Create Pivot Table based on Species and Sepal.Length.Cate.

14.Titanic Casualties – Use the standard 'Titanic' dataset which is part of R Base to answer the following questions.
(i). Use an appropriate apply function to get the sum of males vs females aboard.
(ii). Get a table with the sum of survivors vs sex.
(iii). Get a table with the sum of passengers by sex vs age

DAY 3 – LAB ASSESSMENT

Reg No:
Name:

1. (i) Write a function in R programming to print generate Fibonacci sequence using Recursion in R
.

(ii) Find sum of natural numbers up-to 10, without formula using loop statement.

(iii) create a vector 1:10 and Find a square of each number and store that in a separate list.

2.   (motor trend car road test) comprises fuel consumption, performance and 10 aspects of automobile
design for 32 automobiles. It comes pre-installed with package in R.
 (i)Find the dimension of the dataset
(ii)Give the statistical summary of the features.
(iii)Print the categorical features in Dataset
(iv)Find the average weight(wt) grouped by Engine shape(vs)
(v)Find the largest and smallest value of the variable weight with respect to Engine shape

3.Use ggplot package to plot below EDA questions label the plot accordingly
(i)Create weight(wt) vs displacement(disp) scatter plot factor by Engine Shape(vs)
(ii) Create horsepower(hp) vs mileage (mgp) scatter plot factor by Engine Shape(vs)
(iv)In above plot , Separate columns according to cylinders(cyl) size
(v) Create histogram plot for horsepower (hp) with bin-width size of 5

4. Performing Logistic regression on dataset to predict the cars Engine shape(vs) .
(i)Do the EDA analysis and find the features which is impact the Engine shape and use this for model.
(ii) Split the data set randomly with 80:20 ration to create train and test dataset and create logistic model
(iii)Create the Confusion matrix among prediction and test data.

5. (I) Write R Program to create 15 x15 matrix filled with random numbers between -10 to 10, numbers can repeat. set random seed value to 328

  (ii)Write R Program to display Lower Diagonal and upper Diagonal matrix

  (iii)Write R Program to count 0's in the matrix and check the matrix is sparse matrix or not

  (iv) Write R code to remove outliers. Here the outliers are negative numbers. replace the negative values with positive values

  (v) Find the mean median and mode of the values corresponding to column

  (vi)Find the mean median and mode of the values corresponding to row

DAY 4 – LAB ASSESSMENT

Reg No:

Name:

1.Randomly Sample the iris dataset such as 80% data for training and 20% for test and create Logistics regression with train data, use species as target and petals width and length as feature variables , Predict the probability of the model using test data, Create Confusion matrix for above test model

2. (i)Write suitable R code to compute the mean, median ,mode of the following values
c(90, 50, 70, 80, 70, 60, 20, 30, 80, 90, 20)
(ii) Write R code to find 2nd highest and 3rd Lowest value of above problem.

3. Explore the airquality dataset. It contains daily air quality measurements from New York during a period of five months:
• Ozone: mean ozone concentration (ppb), • Solar.R: solar radiation (Langley),
• Wind: average wind speed (mph), • Temp: maximum daily temperature in degrees Fahrenheit,
• Month: numeric month (May=5, June=6, and so on),• Day: numeric day of the month (1 -4)
 i. Compute the mean temperature(don't use build in function)
ii.Extract the first five rows from airquality.
iii.Extract all columns from airquality except Temp and Wind
iv.Which was the coldest day during the period?
v.How many days was the wind speed greater than 17 mph?

25

4. (i)Get the Summary Statistics of air quality dataset
     (ii)Melt airquality data set and display as a long – format data?
     (iii)Melt airquality data and specify month and day to be "ID variables"?
     (iv)Cast the molten airquality data set with respect to month and date features
     (v) Use cast function appropriately and compute the average of Ozone, Solar.R , Wind and temperature per month?

5.(i) Find any missing values(na) in features and drop the missing values if its less than 10% else replace that with mean of that feature.
    (ii) Apply a linear regression algorithm using Least Squares Method on "Ozone" and "Solar.R"
    (iii)Plot Scatter plot between Ozone and Solar and add regression line created by above model

6. Load dataset named ChickWeight,
    (i).Order the data frame, in ascending order by feature name "weight" grouped by
  feature

"diet" and Extract the last 6 records from order data frame.

(ii).a Perform melting function based on "Chick", "Time", "Diet"   features as ID variables

b. Perform cast function to display the mean value of weight grouped by Diet

c. Perform cast function to display the mode of weight grouped by Diet

7. a. Create Box plot for "weight" grouped by "Diet"

b. Create a Histogram for "weight" features belong to Diet- 1 category

c. Create Scatter plot for " weight" vs "Time" grouped by Diet

8. a. Create multi regression model to find a weight of the chicken , by "Time" and "Diet"
as as
        predictor variables
    b. Predict weight for Time=10 and Diet=1
    c. Find the error in model for same

9 .For this exercise, use the (built-in) dataset Titanic.
    a. Draw a Bar chart to show details of "Survived" on the Titanic based on passenger

Class
    b. Modify the above plot based on gender of people who survived
    c. Draw histogram plot to show distribution of feature "Age"

10. Explore the USArrests dataset, contains the number of arrests for murder, assault, and rape for each of the 50 states in 1973. It also contains the percentage of people in the state who live in an urban area.
   (i) a. Explore the summary of Data set, like number of Features and its type. Find the number   of records for each feature. Print the statistical feature of data
      b. Print the state which saw the largest total number of rape
      c. Print the states with the max & min crime rates for murder
   (ii).a. Find the correlation among the features
      b. Print the states which have assault arrests more than median of the country
      c. Print the states are in the bottom 25% of murder
   (iii). a. Create a histogram and density plot of murder arrests by US stat
        b. Create the plot that shows the relationship between murder arrest rate and proportion
          of the population that is urbanised by state. Then enrich the chart by adding assault
          arrest rates (by colouring the points from blue (low) to red (high)).
      c. Draw a bar graph to show the murder rate for each of the 50 states .

Set 2

1.(i) Write a R program to extract the five of the levels of factor created from a random sample from the LETTERS (Part of the base R distribution.)

 (ii)Write R function to find the range of given vector. Range=Max-Min
Sample input, C<-(9,8,7,6,5,4,3,2,1),
output=8

 (iii)Wirte the R function to find the number of vowels in given string
 Sample input c<- "matrix", output<-2

2.Load inbuild dataset "ChickWeight" in R
(i) Explore the summary of Data set, like number of Features and its type. Fins the number of

records for each features

(ii)Extract last 6 records of dataset

(iii) order the data frame, in ascending order by feature name "weight" grouped by feature "diet"

(iv)Perform melting function based on "Chick","Time","Diet" features as ID variables

(v)Perform cast function to display the mean value of weight grouped by Diet

3.(i)Get the Statistical Summary of "ChickWeight" dataset

(ii)Create Box plot for "weight" grouped by "Diet"

(iii)Create a Histogram for "Weight" features belong to Diet- 1 category
(iv) Create a Histogram for "Weight" features belong to Diet- 4 category
(v) Create Scatter plot for weight vs Time grouped by Diet

4.(i) Create multi regression model to find a weight of the chicken , by "Time" and "Diet" as
as predictor variables
(ii) Predict weight for Time=10 and Diet=1
(iii)Find the error in model for smae

DAY 5 – LAB ASSESSMENT

Reg No:

Name:

1. Randomly Sample the iris dataset such as 80% data for training and 20% for test and create Logistics regression with train data,use species as target and petals width and
length as feature variables ,Predict the probability of the model using test data, Create Confusion matrix for above test model

2.Load dataset named ChickWeight,
i.Order the data frame, in ascending order by feature name "weight" grouped by   feature "diet" and Extract the last 6 records from order data frame.
 (ii).a Perform melting function based on "Chick", "Time", "Diet"   features as ID variables
b. Perform cast function to display the mean value of weight grouped by Diet
c. Perform cast function to display the mode of weight grouped by Diet
(iii)a. Create Box plot for "weight" grouped by "Diet"
b. Create a Histogram for "weight" features belong to Diet- 1 category
c. Create Scatter plot for " weight" vs "Time" grouped by Diet
(iv) a. Create multi regression model to find a weight of the chicken , by "Time" and "Diet" as as predictor variables
        b. Predict weight for Time=10 and Diet=1
         c. Find the error in model for same

3.Explore the USArrests dataset, contains the number of arrests for murder, assault, and rape for each of the 50 states in 1973. It also contains the percentage of people in the state who live in an urban area.
(i) a. Explore the summary of Data set, like number of Features and its type. Find the number of records for each feature. Print the statistical feature of data
b. Print the state which saw the largest total number of rape
c. Print the states with the max & min crime rates for murder

(ii). a. Find the correlation among the features

b. Print the states which have assault arrests more than median of the country

c. Print the states are in the bottom 25% of murder

(iii). a. Create a histogram and density plot of murder arrests by US stat

b. Create the plot that shows the relationship between murder arrest rate and proportion of the population that is urbanised by state. Then enrich the chart by adding assault arrest rates (by colouring the points from blue (low) to red (high)).

c. Draw a bar graph to show the murder rate for each of the 50 states .

4. Using linear regression analysis establish a relationship between height and weight of a person using the input vector given below.

Values of height

151, 174, 138, 186, 128, 136, 179, 163, 152, 131

Values of weight.

63, 81, 56, 91, 47, 57, 76, 72, 62, 48

Predict the weight of a person with height 170 and Visualize the regression graphically

5.For this exercise, use the (built-in) dataset Titanic.
a. Draw a Bar chart to show details of "Survived" on the Titanic based on passenger Class
b. Modify the above plot based on gender of people who survived
c. Draw histogram plot to show distribution of feature "Age"

6 a. Create a data frame based on below table. Month

Month	1	2	3	4	5	6	7	8	9	10	11	12
Spends	1000	4000	5000	4500	3000	4000	9000	11000	15000	12000	7000	3000

Sales	9914	40487	54324	50044	34719	42551	94871	118914	158484	131348	78504	36284

b. Create a regression model for that data frame table to show the amount of sales(Sales) based on the how much the company spends (Spends) in advertising

c. Predict the Sales if Spend=13500

7.a. Create a 6×10 matrix of random integers chosen in the range of from 1:10

b. Find the number of entries in each row which are greater than 4.

c. Which rows contain exactly two occurrences of the number 7?

8.Suppose you track your commute times for two weeks (10 days) and you find the following times in minutes 17 16 20 24 22 15 21 15 17 22 Enter this into R

a. create function "maxi" to find the longest commute time, the function "avger" to find the

average and the function "mini" to find the minimum.
b. Oops, the 24 was a mistake. It should have been 18. How can you fix this? Do so, and then find the new average.
c. How many times was your commute 20 minutes or more?

9.a. Create a 3x4 matrix with 12 random numbers between 1-100; have the matrix be filled our row-by-row, instead of column-by-column. Name the columns of the matrix uno, dos, tres, cuatro, and the rows x, y, z. Scale the matrix by 10 and save the result.
b. Extract the column called "uno" as a vector from the original matrix and save the result
c. Extract the row called 'y' as a vector from the original matrix and print thesum of the vector.

10. a. Write suitable R code to compute the mean, median ,mode of the following values
 c(90, 50, 70, 80, 70, 60, 20, 30, 80, 90, 20,75,70,10,60,70,85,95,55,15)
 b. Write R code to find 2nd highest and 4th Lowest value of above problem.

11. a. Write a program for creating a pie-chart in R using the input vector (21,62,10,53). Provide labels for the chart as 'London', 'New York', 'Singapore', 'Mumbai'. Add a title to the chart as 'city pie-chart' and add a legend at the top right corner of the chart.
b. Write a program for creating a bar chart using the vectors H=c(7,12,28,3,41) and M=c("mar", "apr", "may", "jun", "jul"). Add a title to the chart as "Revenue chart"

c. Creat histogram plot for data below with suitable bin-width
v <- c(19, 23, 11, 5, 16, 21, 32, 14, 19, 27, 39, 120, 40, 70, 90)