AIT362	PROGRAMMING IN R	CATEGORY	L	T	P	CREDIT	YEAR OF INTRODUCTION
		PEC	2	1	0	3	2019

Preamble: The objective of this course is to enable the learner to make use of R Programming language to perform analysis and extraction of information from data irrespective of the quantity. It encompasses the R programming environment, syntax, data representations, data processing, statistical analysis and visualization. This course facilitates the learner to develop modular software solutions to perform statistical analysis and data extraction.

Prerequisite: Fundamental concepts in programming in C and Probability and Statistical Modeling

Course Outcomes: After the completion of the course the student will be able to:

CO 1	Illustrate uses of conditional and iterative statements in R programs. (Cognitive Knowledge level: Apply)
CO 2	Write, test and debug R programs (Cognitive Knowledge level: Apply)
CO 3	Illustrate the use of Probability distributions and basic statistical functions. (Cognitive Knowledge level: Apply)
CO 4	Visualize different types of data (Cognitive Knowledge level: Apply)
CO 5	Comprehend regression modeling using R (Cognitive Knowledge level: Understand)

Mapping of course outcomes with program outcomes

	PO 1	PO 2	PO 3	PO 4	PO 5	PO6	PO 7	PO 8	PO 9	PO10	PO11	PO12
CO1	✓	✓	✓		✓							✓
CO2	✓	✓	✓		✓							✓
CO3	✓	✓	✓	✓	✓							✓
CO4	✓	✓	✓	✓	✓							✓
CO5	✓	✓			✓							✓

Abstract POs defined by National Board of Accreditation			
PO#	**Broad PO**	**PO#**	**Broad PO**
PO1	Engineering Knowledge	**PO7**	Environment and Sustainability
PO2	Problem Analysis	**PO8**	Ethics
PO3	Design/Development of solutions	**PO9**	Individual and team work
PO4	Conduct investigations of complex problems	**PO10**	Communication
PO5	Modern tool usage	**PO11**	Project Management and Finance
PO6	The Engineer and Society	**PO12**	Life long learning

Assessment Pattern

Bloom's Category	**Continuous Assessment Tests**		**End Semester Examination Marks**
	Test1 (percentage)	**Test2 (percentage)**	
Remember	20	20	20
Understand	40	40	40
Apply	40	40	40
Analyze			
Evaluate			
Create			

Mark distribution

Total Marks	**CIE Marks**	**ESE Marks**	**ESE Duration**
150	50	100	3 hours

Continuous Internal Evaluation Pattern:

Attendance: 10 marks
Continuous Assessment Tests : 25 marks
Continuous Assessment Assignment: 15 marks

Internal Examination Pattern:
Each of the two internal examinations has to be conducted out of 50 marks

First Internal Examination shall be preferably conducted after completing the first half of the syllabus and the Second Internal Examination shall be preferably conducted after completing the remaining part of the syllabus.

There will be two parts: Part A and Part B. Part A contains 5 questions (preferably, 2 questions each from the completed modules and 1 question from the partly covered module), having 3 marks for each question adding up to 15 marks for part A. Students should answer all questions from Part A. Part B contains 7 questions (preferably, 3 questions each from the completed modules and 1 question from the partly covered module), each with 7 marks. Out of the 7 questions in Part B, a student should answer any 5.

End Semester Examination Pattern:

There will be two parts; Part A and Part B. Part A contains 10 questions with 2 questions from each module, having 3 marks for each question. Students should answer all questions. Part B contains 2 questions from each module of which a student should answer any one. Each question can have a maximum of 2 subdivisions and carries 14 marks.

SYLLABUS

Module -1 (Introduction to R)
The R Environment - Command Line Interface and Batch processing, R Packages, Variables, Data Types, Vectors- vector operations and factor vectors, List- operations, Data Frames, Matrices and arrays, Control Statements- Branching and looping - For loops, While loops, Controlling loops. Functions- Function as arguments, Named arguments

Module -2(Reading and writing data)
Importing data from Text files and other software, Exporting data, importing data from databases- Database Connection packages, Missing Data - NA, NULL
Combining data sets, Transformations, Binning Data, Subsets, summarizing functions. Data Cleaning, Finding and removing Duplicates, Sorting.

Module -3 (Statistics with R)
Analyzing Data, Summary statistics, Statistical Tests- Continuous Data, Discrete Data, Power tests, Common distributions- type arguments. Probability distributions, Normal distributions

Module -4(Data Visualization)
R Graphics- Overview, Customizing Charts, Graphical parameters, Basic Graphics functions, Lattice Graphics - Lattice functions, Customizing Lattice Graphics, Ggplot.

Module - 5 (Regression Models)
Building linear models - model fitting, Predict values using models, Analyzing the fit, Refining the model, Regression- types, Unusual observation and corrective measures,

Comparison of models, Generalized linear models - Logistic Regression, Poisson Regression, Nonlinear least squares

Text Book

1. Joseph Adler, " R in a Nutshell", Second edition,O'reilly,2012

Reference Books

1. Jared P Lander, R for Everyone- Advanced analytics and graphics, Addison Wesley data analytics series, Pearson
2. Norman matloff, The art of R programming, A Tour of Statistical, Software Design, O'reilly
3. Robert Kabacoff, R in action, Data analysis and graphics with R, Manning
4. Garret Grolemund, Hands-on programming with R, Write your own functions and simulations, O'reilly

Sample Course Level Assessment Questions

Course Outcome 1 (CO1):

1. What is Coercion? How is it done in R?
2. Write a program to find the factorial of a number.
3. Write a program to compute roots of a quadratic equation.

Course Outcome 2 (CO2):

1. Write a program to read data from a table 'table123' in a database named 'db123' and display the values .
2. Explain Data cleaning in R
3. How missing data is handled in R?

Course Outcome 3(CO3):

1. Explain summary function in R
2. Illustrate how statistical testing is performed in R
3. Describe about probability distributions.

Course Outcome 4 (CO4):

1. Illustrate the use of ggplot() and various data visualization tools using appropriate datasets

Course Outcome 5 (CO5):

1. Illustrate the steps to predict the weight of a person when his height is unknown using linear regression for the data given below.

Height	151	174	138	186	128	136	179	163	152	130
Weight	63	81	56	91	47	57	76	72	62	48

Model Question Paper

QP CODE: PAGES:3

Reg No:_____

Name :_____

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

SIXTH SEMESTER B.TECH DEGREE EXAMINATION, MONTH & YEAR

Course Code: **AIT 362**

Course Name: Programming in R

Max.Marks:100 **Duration: 3 Hours**

PART A

Answer all Questions. Each question carries 3 Marks

1. Write a R program to add element "23" to the vector (24,56,67) in the second position.
2. Discuss the general list operations in R with example.
3. Calculate the cumulative sum and cumulative product for the given data 23, 1, 7,2,8,10, 17 using R Program.
4. Explain aggregate function in R.
5. List the applications of R programming.
6. Illustrate summary function.
7. List any three graphics functions.
8. Explain Lattice function.
9. Suppose that you have a dataset D1 and you design a linear regression model of degree 3 polynomial and you found that the training and testing error is "0" or in other terms it perfectly fits the data. What will happen when you fit a degree 2 polynomial in linear regression?
10. Explain logistic regression function in R.

(10x3=30)

Part B

Answer any one Question from each module. Each question carries 14 Marks

11.a Write a R program to extract every nth element from a vector. (7 marks)

11.b Find the Nth highest value of a vector in R. (7 marks)

OR

12.a Write a R program to create a data frame using two given vectors and display the duplicate elements and unique rows of the said data frame. (7 marks)

12.b Write a R program to compare two data frames to find the row(s) in the first data frame that are not present in the second data frame. (7 marks)

13.a Write a R program to call the (built-in) dataset air quality. Remove the variables 'Solar.R' and 'Wind' and display the data frame. (7 marks)

13.b Illustrate transformation functions in R. (7 marks)

OR

14.a Write a R program to write the following data to a CSV file. (7 marks)

	Country	Population_1_july_2018	Population_1_july_2019	change_in_percents
1	China	1,427,647,786	1,433,783,686	+0.43%
2	India	1,352,642,280	1,366,417,754	+1.02%
3	United States	327,096,265	329,064,917	+0.60%
4	Indonesia	267,670,543	270,625,568	+1.10%
5	Pakistan	212,228,286	216,565,318	+2.04%

14.b Given a file "auto.csv" of automobile data with the fields index, company, body-style, wheel-base, length, engine-type, num-of-cylinders, horsepower, average-mileage, and price, write R program to print total cars of all companies, Find the average mileage of all companies. (7 marks)

15.a Write a note on data analysis using R. (7 marks)

15.b Explain how statistical test are performed using R functions. (7 marks)

OR

16.a Write R code to generate the probability distribution table for number of successes from a binomial distribution where n=5 and probability of success in each trial is 0.25. (7 marks)

16.b Fit a Poisson distribution with the following data using the following data (7 marks)

X	0	1	2	3	4	5
F	142	156	69	27	5	1

OR

17 Given the sales information of a company as CSV file with the following, fields month_number, face cream, facewash, toothpaste, bathingsoap, shampoo, moisturizer, total_units, total_profit. Write R codes to visualize the data as follows:

a) Toothpaste sales data of each month and show it using a scatter plot. (7 marks)

b) Calculate total sale data for last year for each product and show it using a Pie chart. (7 marks)

OR

18.a Explain ggplot() with and example. (7 marks)

18.b Describe how categorical data is visualized using R. (7 marks)

19.a Illustrate model fitting in simple linear model. (7 marks)

19.b Explain different types of regression. (7 marks)

OR

20.a Describe the unusual observations in the regression model. (7 marks)

20.b Explain corrective measures of unusual observations in regression modelling. (7 marks)

TEACHING PLAN

No	Contents	No of Lecture Hours (35 Hours)
	Module -1 (Introduction to R)	**(8 hours)**
1.1	The R Environment- Command Line Interface and Batch processing, R Packages	1 hour
1.2	Variables, Data Types	1 hour
1.3	Vectors- vector operations and factor vectors	1 hour
1.4	List- List operations, Data Frames	1 hour
1.5	Matrices and arrays	1 hour
1.6	Control Statements- If and else, switch, if else	1 hour
1.7	Loops- For loops, While loops, Controlling loops	1 hour
1.8	Functions- Function as arguments, Named arguments	1 hour
	Module -2(Reading and writing data)	**(8 hours)**
2.1	Importing data from Text files and other software, Exporting data	1 hour
2.2	Importing data from databases- Database Connection packages	1 hour
2.3	Missing Data-NA, NULL	1 hour
2.4	Combining data sets, Transformations	1 hour
2.5	Binning Data, Subsets, summarizing functions	1 hour
2.6	Data Cleaning	1 hour
2.7	Finding and removing Duplicate	1 hour
2.8	Sorting	1 hour
	Module -3 (Statistics with R)	**(6 hours)**
3.1	Analyzing Data	1 hour
3.2	Summary statistics	1 hour
3.3	Statistical Tests- Continuous Data, Discrete Data, Power tests	1 hour
3.4	Common distributions- type arguments	1 hour
3.5	Probability distributions	1 hour
3.6	Normal distributions	1 hour
	Module -4(Data Visualization)	**(6 hours)**
4.1	R Graphics- Overview	1 hour
4.2	Customizing Charts	1 hour
4.3	Graphical parameters, Basic Graphics functions	1 hour
4.4	Lattice Graphics - Lattice functions	1 hour
4.5	Customizing Lattice Graphics	1 hour
4.6	ggplot	1 hour
	Module - 5 (Regression Models)	**(7 hours)**

5.1	Building linear models - model fitting	1 hour
5.2	Predict values using models, Analyzing the fit, Refining the model	1 hour
5.3	Regression- types of regression	1 hour
5.4	Unusual observations and corrective measures	1 hour
5.5	Comparison of models	1 hour
5.6	Generalized linear models -Logistic Regression, Poisson Regression	1 hour
5.7	Nonlinear least squares	1 hour