Rayat Shikshan Sanstha's

Karmaveer Bhaurao Patil College, Vashi

Autonomous

Affiliated to University of Mumbai

Syllabus

Sr. No.	Heading	Particulars
1	Title of Course	Master in Data Science
2	Eligibility for Admission	Students with a Bachelor's degree in Mathematics / Statistics / Computer Science /Computer Application/ Information Technology / Physics/B.E. in Computer Science / Information Technology from a recognized university with a minimum aggregate score of 50% or higher are eligible for this course.
3	Passing marks	40%
4	Ordinances/Regulations (if any)	
5	Duration	Course Duration of Master of Science [M.Sc] (Data Science) is 2 Years.
6	Level	P.G.
7	Pattern	Choice Based Credit, Grading and Semester
8	Status	New
9	To be implemented from	2021-2022
	Academic year	

AC - 25/10/2021
Item No - 5.5

Rayat Shikshan Sanstha's
KARMAVEER BHAURAO PATIL COLLEGE, VASHI.
NAVI MUMBAI
(AUTONOMOUS COLLEGE)
Sector-15- A, Vashi, Navi Mumbai - 400 703

Syllabus for M.Sc. in Data Science

Program: M.Sc. Data Science I

(Choice Based Credit, Grading and Semester System with effect from the academic year 2021-22)

Preamble

This syllabus is an honest attempt to include following ideas, among other things, into practice:

- Create a unique identity for MSc in Data Science distinct from similar degrees in other related subjects.
- Recommend provision for specialization in Data Science.
- Specialized knowledge of the central concepts, theories, and research methods of data science as well as applied skills.
- Specialized knowledge of computer science theories, methods, practices and strategy.
- Understanding of statistical, mathematical concepts in the context of data science.
- Understanding of various analysis tools and software used in data science.
- Awareness of rapid technological changes.
- Analytical and critical thinking skills.
- Written and oral communication skills, including presentations and report writing.

M.Sc in Data Science is a postgraduate course that comes under the set of science as a major field of study. The duration of the course is 2 years which is equally divided into 4 semesters.In the third semester,one of the course is an internship. M.Sc Data Science course syllabus is designed in a manner that covers all the aspects of Data Science.

The syllabus proposes to have three core compulsory courses, one Skill Enhancement Course and one Discipline Specific Elective course in semester I. Semester II also proposes three core compulsory courses, one Skill Enhancement Course and one Discipline Specific Elective course. The course gives insights into the practical and theoretical aspects of data science, Big data analytics, Business Analytics, Real-Time Processing, Neural Networks, Artificial Intelligence, and Machine Learning. The primary focus of the course is to equip the candidates of the course with principal concepts of data science and application of the same in real-time processing and applications.

Data science combines the knowledge of mathematics, computer science and statistics to solve exciting data-intensive problems in industry and in many fields of science. As data is collected and analysed in all areas of society, demand for professional data scientists is high and will grow higher.

We thank all the industry experts, senior faculties and our colleagues of different colleges as well as BOS members who have given their valuable comments and suggestions, which we tried to incorporate.

Choice Based Credit Semester System
Academic year 2021-2022

SEMESTER - I

CODE	COURSE TYPE	SUBJECT	SCHEME OF INSTRUCTION (PERIOD PER WEEK)		SCHEME OF EXAMINATION (MAX MARKS)			NO. OF CREDITS
			TH	LAB	CIA	SEE	TOTAL	
PGDS101	CORE	Advance Database Technologies	4	-	40	60	100	4
PGDS102	CORE	Descriptive Statistics and Probability	4	-	40	60	100	4
PGDS103	CORE	Applied Linear Algebra	4	-	40	60	100	4
PGDS104 / PGCS 104	Skill Enhancement Elective-I	Data Visualization using R/ Advanced Python programming	3	-	40	60	100	3
PGDS105	Discipline Specific Elective-I **OR**	Data Warehousing & Mining	4	-	40	60	100	4
PGDS106	Discipline Specific Elective-II	Data Structure with Python	4	-	40	60	100	4
PGDSP101	Core Subject Practical	PRACTICAL ON ADVANCE DATABASE TECHNOLOGIES PGDS101	-	4			50	2
PGDSP102	Core Subject Practical	PRACTICAL ON DESCRIPTIVE STATISTICS AND PROBABILITY PGDS102	-	4			50	2
PGDSP103	Core Subject Practical	PRACTICAL ON APPLIED LINEAR ALGEBRA PGDS103	-	4			50	2
PGDSP104 / PGCSP104	Skill Enhancement Practical	PRACTICAL ON ADVANCED PYTHON PROGRAMMING PGDS104/PGCS104	-	2			50	1
PGDSP105A	Discipline Specific Elective-I Practical **OR**	PRACTICAL ON DATA WAREHOUSING & MINING PGDS105A	-	4			50	2
PGDSP105B	Discipline Specific Elective-I Practical	PRACTICAL ON DATA STRUCTURE WITH PYTHON PGDS105B	-	4			50	2
						TOTAL	750	28

SEMESTER - II

CODE	COURSE TYPE	SUBJECT	SCHEME OF INSTRUCTION	SCHEME OF EXAMINATION	NO. OF CREDITS

			(PERIOD PER WEEK)		(MAX MARKS)			
			TH	LAB	CIA	SEE	TOTAL	
PGDS201	CORE	Research in Computing	4	-	40	60	100	4
PGDS202	CORE	Optimization Techniques	4	-	40	60	100	4
PGDS203	CORE	Statistical Inference	4	-	40	60	100	4
PGDS204	Skill Enhancement Elective-I	Advanced Python Programming	2	-		60	100	2
PGDS205	Discipline Specific Elective-I OR	Big Data Analytics	4	-	40	60	100	4
PGDS206	Discipline Specific Elective-I	Analysis of Algorithm	4	-	40	60	100	4
PGDSP201	Core Subject Practical	PRACTICAL ON RESEARCH IN COMPUTING PGDS201	-	4			50	2
PGDSP202	Core Subject Practical	PRACTICAL ON OPTIMIZATION TECHNIQUES PGDS202	-	4			50	2
PGDSP203	Core Subject Practical	PRACTICAL ON STATISTICAL INFERENCE PGDS203	-	4			50	2
PGDSP204	Skill Enhancement Practical	PRACTICAL ON ADVANCED PYTHON PROGRAMMING PGDS204	-	4			50	2
PGDSP205A	Discipline Specific Elective-I Practical	PRACTICAL ON BIG DATA ANALYTICS PGDS205A	-	4			50	2
PGDSP205B	Discipline Specific Elective-I Practical	PRACTICAL ON ANALYSIS OF ALGORITHM PGDS205B	-	4			50	2
						TOTAL	750	28

Note: TH-Theory, CIA- Continuous Internal Assessment, SEE-Semester End Examination.

Semester I – Theory

Class: M.Sc	**Branch**: Data Science	**Semester:** I
Subject: Advanced Database Technologies		

Period per Week(Each 60 mins)	Lecture	04	
	Practical	04	
Evaluation System		**Hours**	**Marks**
	Semester End Exam	2 hrs.30min	60
	Continuous Internal Assessment	—	40
	Semester End Practical Examination	2 hrs.	50
	Total	—	150

Course: PGDS101	**Advanced Database Technologies** (Credits : 4 Lectures/Week: 4)	**Lectures**
	Expected Learning Outcomes: After successful completion of this course, students would be able to 1 .Recall the concept of Database Systems, Relational Databases ,Structure of Relational Databases & Relational Algebra. 2.Describe the Object Databases Systems ,Design the E-R model ,Normalization process. 3. Illustrate the NOSQL concept with the NOSQL database. 4. Explain Data Modeling With Graph (NEeo4j), Key-Value Databases (Riak), Column-Family stores (Cassandra).	
Unit I	**Introduction:** Purpose of Database Systems, View of Data:Data Abstraction,Instance and Schemas, Relational Databases: Tables, DML, DDL, Data storage and querying: Storage Manager, The query processor , Database Architecture , Speciality Databases **Introduction to Relational Model :** Structure of Relational Databases ,Database Schema , Keys ,Relational Algebra	**8**
Unit II	**Database Design and E-R model** :Overview of the Design process and Entity Relationship ,Functional Dependency, Anomalies in a Databases,Normalization process: Conversion to first normal form, Conversion to second normal form, Conversion to third normal form, The Boyce-Codd Normal Form (BCNF), Fourth Normal form and fifth normal form Denormalization, **Object Databases Systems:** Overview of Object -Oriented concepts & characteristics Objects,,OIDs and reference types, Database design for ORDBMS,Comparing RDBMS, OODBMS & ORDBMS	**10**
Unit III	**Introduction to NOSQL (Core concepts):** Why NoSQL,Brief History of NoSQL Databases ,Features of NoSQL,Types of NoSQL Databases,CAP Theorem,Aggregate Data Models, Data modeling details ,Distribution Models, Consistency ,Version stamps, Map-Reduce **Implementation with NOSQL databases:** Document Databases (Mongodb) MongoDB Features, MongoDB Example,Key Components of MongoDB Architecture,Why Use MongoDB,Data Modelling in MongoDB,Difference between MongoDB & RDBMS	**15**

Unit IV	**Data Modeling With Graph (NEeo4j):** Comparison of Relational and Graph Modeling, Property Graph Model Graph Analytics: Link analysis algorithm- Web as a graph, Page Rank- Markov chain, page rank computation, Topic specific page rank (Page Ranking Computation techniques: iterative processing, Random walk distribution Querying Graphs: Introduction to Cypher, case study: Building a Graph Database Application- community detection. **Key-Value Databases (Riak):** From array to key value databases, Essential features of key value Databases, Properties of keys, Characteristics of Values, Key-Value Database Data Modeling Terms, Key-Value Architecture and implementation Terms, Designing Structured Values, Limitations of KeyValue Databases, Design Patterns for Key-Value Databases, Case Study: Key-Value Databases for Mobile Application Configuration	**15**
Unit V	**Column-Family stores (Cassandra)** Data warehousing schemas: Comparison of columnar and row-oriented storage, Column-store Architectures: C-Store and Vector-Wise, Column-store internals and, Inserts/updates/deletes, Indexing, Adaptive Indexing and Database Cracking Advanced techniques: Vectorized Processing, Compression, Write penalty, Operating Directly on Compressed Data Late Materialization Joins , Group-by, Aggregation and Arithmetic Operations, Case Studies	**12**

Text book:

- NoSQL Distilled Pramod Sadalge, Martin Fowler
- Next Generation database: NoSQL and big data by Guy Harrison

Reference

- NoSQL for Dummies A Willy Brand

Links:

- **https://hostingdata.co.uk/nosql-database/**
- **https://www.guru99.com/what-is-mongodb.html**

Sr. No.	Practicals of PGDSP101
1	Practical on Relational Algebra, SQL Commands,Normalization
2	How to Download & Install MongoDB on Windows
3	Hello World MongoDB: JavaScript Driver
4	● Install Python Driver ● Install Ruby Driver
5	Install MongoDB Compass- MongoDB Management Tool MongoDB Configuration, Import, and Export

6	Download a zip code dataset at http://media.mongodb.org/zips.json .Use mongo import to import the zip code dataset into MongoDB. After importing the data, answer the following questions by using aggregation pipelines: (1) Find all the states that have a city called ”BOSTON”. Find all the states and cities whose names include the string ”BOST”. Each city has several zip codes. Find the city in each state with the most number of zip codes and rank those cities along with the states using the city populations. MongoDB can query on spatial information.
7	Master Data Management using Neo4j Manage your master data more effectively The world of master data is changing. Data architects and application developers are swapping their relational databases with graph databases to store their master data. This switch enables them to use a data store optimized to discover new insights in existing data,provide 360-degree view of master data and answer questions about data relationships in real time
8	Create a database that stores road cars. Cars have a manufacturer ,a type. Each car has a maximum performance and a maximum torque value. Do the following: Test Cassandras replication schema and consistency models.
9	Case Study

Class: M.Sc-I	**Branch**: Data Science	**Semester:** I	
Subject: Descriptive Statistics and Probability			
Period per Week(Each 60 mins)	Lecture	04	
	Practical	04	
Evaluation System		**Hours**	**Marks**
	Semester End Exam	2 hrs.30min	60
	Continuous Internal Assessment	__	40
	Semester End Practical Examination	2 hrs.	50
	Total	__	150

Course: PGDS102	Descriptive Statistics and Probability **(Credits : 4 Lectures/Week: 4)**	**Lectures**
	Expected Learning Outcomes: After successful completion of this course, students would be able to 1. Describe the data and its properties by use of central tendency and variability. 2. Explain the concepts of probability and its distributions. 3. Apply sampling distributions to contribute to the process of making rational decisions in analytical problems 4. Analyze the relationship between two quantitative variables using Correlation and Regression	

Unit I	**Descriptive Statistics and Introduction to Probability:** Measures of Central Tendency: Mean, Median, Mode Partition Values: Quartiles, Percentiles, Box Plot Measures of Dispersion: Variance, Standard Deviation, Coefficient of variation Skewness: Concept of skewness, measures of skewness Kurtosis: Concept of Kurtosis, Measures of Kurtosis. Probability - classical definition, probability models, axioms of probability, probability of an event. **Concepts and definitions of conditional probability, multiplication theorem $P(A \cap B) = P(A).P(B|A)$ Bayes' theorem (without proof)** Concept of Posterior probability, problems on posterior probability. Definition of sensitivity of a procedure, specificity of a procedure. Application of Bayes' theorem to design a procedure for false positive and false negative. Concept and definition of independence of two events. Numerical problems related to real life situations.	**15 L**
Unit II	**Introduction to Random Variables** Definition of discrete random and continuous random variable. Concept of Discrete and Continuous probability distributions. (p.m.f. and p.d.f.). Distribution function, Expectation and variance, Numerical problems related to real life situations.	**15 L**
Unit III	**Special Distributions** Binomial Distribution, Uniform Distribution, Poisson Distribution, Negative Binomial Distribution, Geometric Distribution, Continuous Uniform, Distribution, Exponential Distribution, Normal Distribution, Log Normal Distribution, Gamma Distribution, Weibull Distribution, Pareto Distribution. (For all the probability distributions its pmf/pdf, p-p plot, q-q plot, generation of probabilities and random samples using R software is expected.)	**15 L**
Unit IV	**Correlation and Regression** Bivariate data, Scatter diagram. Correlation, Positive Correlation, Negative correlation, Zero Correlation, Karl Pearson's coefficient of correlation (r), limits of r ($-1 \leq r \leq 1$), interpretation of r, Coefficient of determination (r^2), Meaning of regression, difference between correlation and regression. Fitting of line $Y = a+bX$, Concept of residual plot and mean residual sum of squares. Multiple correlation coefficient, concept, definition, computation and interpretation. Partial correlation coefficient, concept, definition, computation and interpretation. Multiple regression plane. Identification and solution to Multicollinearity. Evaluation of the Model using R square and Adjusted R square. Introduction to logistic regression, Difference between linear and logistic regression, Logistic equation, How to build logistic regression model in R, Odds ratio in logistic regression.	**15 L**

	All topics to be covered for raw data using R software. Manual calculations are not expected.	

Text book:

- Fundamentals of Applied Statistics (3rd Edition), Gupta and Kapoor, S.Chand and Sons, New Delhi, 1987.
- An Introductory Statistics, Kennedy and Gentle.

Reference

1. Statistical Methods, G.W. Snedecor, W.G. Cochran, John Wiley & sons, 1989.
2. Introduction to Linear Regression Analysis, Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining, Wiley.
3. Modern Elementary Statistics, Freund J.E., Pearson Publication, 2005.
4. Probability, Statistics, Design of Experiments and Queuing theory with applications Computer Science, Trivedi K.S., Prentice Hall of India, New Delhi,2001.
5. A First course in Probability 6th Edition, Ross, Pearson Publication, 2006.
6. Introduction to Discrete Probability and Probability Distributions, Kulkarni M.B., Ghatpande S.B., SIPF Academy, 2007.
7. A Beginners Guide to R, Alain Zuur, Elena Leno, Erik Meesters, Springer, 2009.
8. Statistics Using R, Sudha Purohit, S.D.Gore, Shailaja Deshmukh, Narosa, Publishing Company

Links:

- **https://www.dcpehvpm.org/E-Content/Stat/FUNDAMENTAL%20OF%20MATHEMATICAL%20STATISTICS-S%20C%20GUPTA%20&%20V%20K%20KAPOOR.pdf**
- **https://www.mathsisfun.com/data/random-variables.html**

Sr. No.	Practicals of PGDSP102
1	Introduction to R-studio, mathematical and logical operators in R, Data types and data structures, simple operations and programs, matrix operations
2	Data frames, string operations, factors, handling categorical data, lists and list
3	Operations Loops and conditional statements, switch and break function

4	Apply functions, Statistical problem solving in R,
5	Visualizations in R – 1
6	Visualizations in R – 2
7	Spatial Data Representation and Graph Analysis.
8	Hands-on data manipulations1: cleaning, sub-setting, sampling, data transformations and allied data operations
9	Hands-on data manipulations2: cleaning, sub-setting, sampling, data transformations and allied data operations
10	Case Study

Class: M.Sc	**Branch**: Data Science	**Semester:** I	
Subject: Applied Linear Algebra			
Period per Week(Each 60 min)	Lecture	04	
	Practical	04	
Evaluation System		**Hours**	**Marks**
	Semester End Exam	2 hrs.30min	60
	Continuous Internal Assessment	__	40
	Semester End Practical Examination	2 hrs.	50
	Total	__	150

Course: PGDS103	**Applied Linear Algebra** **(Credits : 3 Lectures/Week: 4)**	**Lectures**
	Expected Learning Outcomes: After successful completion of this course, students would be able to 1. Describe the concept of characteristic polynomial, eigenvalues and eigenvectors. 2. Recognize and use equivalent forms to identify matrices and solve linear systems of equations. 3. Explain how orthogonal projections relate to least square approximations. 4. Acquire the knowledge of various concepts in Applied Algebra. 5. Employ Python to perform various matrix and vector computations.	
Unit I	**Matrices** Matrices: Introduction to Matrices, Zero and identity Matrices, Transpose, addition and Matrix Multiplication, Geometric Transformation, Linear and Orthogonal	**15 L**

	Transformations Rank of matrix, normal form, Consistency, System of Linear Equations, Eigenvalues and eigenvectors.	
Unit II	**Vectors** Vector: Vector addition, Scalar Vector multiplication, unit vector, norm of vector. Linear Functions, Linear Combinations, Linearly dependent and independence, Basis.	**15 L**
Unit III	**Inner Product Space :** Inner Product Spaces, Norms and Distance: Orthogonality Inner products, Cauchy-Schwarz inequality, Orthogonal projections, Gram-Schmidt orthogonalization, Matrix representation of inner product.	**15 L**
Unit IV	**Least Squares** Least Squares: Least Squares Problem, Solution, Solving Least Squares Problems, Examples. **Least squares data fitting:** Least Squares data fitting, Validation, Feature Engineering. **Least Squares Classification:** Classification, Least Squares Classifier, Multiclassifiers **Multi Objective Least Squares:** Multi Objective Least Squares, Control, Estimation and Inversion, Regularised data fitting, Complexity Constrained Least Squares: Constrained Least Squares problem, Solution, Solving constrained Least Squares problems. Constrained Least Squares Applications: Portfolio Optimization, Linear Quadratic control, Linear Quadratic State Estimation.	**15 L**

Textbooks:

1. Advanced Engineering Mathematics by Erwin Kreyszig (Wiley Eastern Ltd.)
2. Introduction to Applied Linear Algebra Vectors, Matrices and Least Squares by Stephen Boyd (Stanford University) and Lieven Vandenberghe (University of California, Los Angeles) Cambridge University Press.

References:

1. Least Squares Regression Analysis in Terms of Linear Algebra By Enders A. Robinson
2. Kenneth H. Rosen's Discrete Mathematics and Its Applications with Combinatorics and Graph Theory 7th Edition(McGraw-Hill Education)
3. Higher Engineering Mathematics by B. S. Grewal (Khanna Publication, Delhi) Reference

Links :

- **https://www.google.co.in/books/edition/Introduction_to_Applied_Linear_Algebra/IApaDwAAQBAJ?hl=en&gbpv=1&dq=Least+Squares+for+algebra&printsec=frontcover**

Sr. No.	**Practicals of PGDSP103**
1	Introduction to numpy and sympy.
2	Write a program to do the following:

	1. Enter a vector u as a n-list 1. Enter another vector v as a n-list 2. Find the vector addition 3. Find the scalar vector multiplication
3	Write a program to do the following: 1. Enter a vector u as a n-list 2. Enter another vector v as a n-list 3. Find the linear Independence & Dependance of vectors
4	Write a program to find the inner product of two vectors.
5	Write a program on The K means algorithm
6	Write a program to do the following: 1. Enter a vector b and find the projection of b orthogonal to a given vector u. 2. Find the projection of b orthogonal to a set of given vectors
7	Write a program to do the following: 1. Enter an r by c matrix M (r and c being positive integers) 2. Display M in matrix format 3. Display the rows and columns of the matrix M 4. Find the scalar multiplication of M for a given scalar. 5. Find the transpose of the matrix M.
8	Write a program to Find the vector –matrix multiplication of a r by c matrix M with an c-vector u
9	Write a program to enter a matrix and check if it is invertible. If the inverse exists, find the inverse.
10	Write a program to solve system of linear equation

Class: M.Sc	**Branch**: Data Science	**Semester:** I	
Subject: Data Visualization using R			
Period per Week(Each 60 min)	Lecture	03	
	Practical	01	
Evaluation System		**Hours**	**Marks**
	Semester End Exam	2 hrs.30min	60
	Continuous Internal Assessment	__	40
	Semester End Practical Examination	2 hrs.	50
	Total	__	150

Course: PGDS104	**Data Visualization using R** **(Credits : 4 Lectures/Week: 3)**	**Lectures**
	Expected Learning Outcomes: After successful completion of this course, students would be able to 1. Explain basic programming language concepts using R	

	2. Differentiate between different R data structures such as: string, number, vector, matrix, data frame, factor, date and time object 3. Collect detailed information raw data using R profiler 4. Visualize your data using base R graphics	
Unit I	**Overview of R :**History and Overview of R- Basic Features of R-Design of the R System- Installation of R- Console and Editor Panes- Comments- Installing and Loading R Packages- Help Files and Function DocumentationSaving Work and Exiting R- Conventions- R for Basic Math- Arithmetic- Logarithms and Exponentials E-Notation- Assigning Objects- Vectors- Creating a Vector-Sequences, Repetition, Sorting, and Lengths- Subsetting and Element Extraction-Vector-Oriented Behaviour	**15 L**
Unit-II	**Matrices And Arrays:** Defining a Matrix – Defining a Matrix- Filling Direction-Row and Column Bindings- Matrix DimensionsSubsetting- Row, Column, and Diagonal Extractions- Omitting and Overwriting- Matrix Operations and Algebra-Matrix Transpose- Identity Matrix- Matrix Addition and Subtraction- Matrix MultiplicationMatrix Inversion-Multidimensional Arrays- Subsets, Extractions, and Replacements	**15 L**
Unit-III	**Non-numeric Values :**Logical Values- Relational Operators- Characters- Creating a String- Concatenation- Escape SequencesSubstrings and Matching- Factors-Identifying Categories- Defining and Ordering Levels- Combining and Cutting Lists And Data Frames:Lists of Objects-Component Access-Naming-Nesting-Data Frames-Adding Data Columns and Combining Data Frames-Logical Record Subsets-SomeSpecial,Values-Infinity-NaN-NA-NULLAttributes-Object-Class-Is-Dot Object-Checking Functions-As-Dot Coercion Functions	**15 L**
Unit-IV	**Basic Plotting:**Using plot with Coordinate Vectors-Graphical Parameters-Automatic Plot Types-Title and Axis LabelsColor-Line and Point Appearances-Plotting Region Limits-Adding Points, Lines, and Text to an Existing Plot-ggplot2 Package-Quick Plot with ggplot-Setting Appearance Constants with Geoms--READING AND WRITING FILES- R-Ready Data Sets- Contributed Data Sets-Reading in External Data Files- Writing Out Data Files and Plots- Ad Hoc Object Read/Write Operations	**15 L**

TextBook:

1. https://www.cs.upc.edu/~robert/teaching/estadistica/rprogramming.pdf
2. Tilman M.Davies,“THE BOOK OF R - A FIRST PROGRAMMING AND STATISTICS” Library of Congress Cataloging-in-Publication Data,2016

References:

1. Wickham, H. & Grolemund, G. (2018). for Data Science. O’Reilly: New York. Available
2. Steven Keller, “R Programming for Beginners”, CreateSpace Independent Publishing Platform 2016
3. Kun Ren ,”Learning R Programming”, Packt Publishing,2016

Links:

- **https://r4ds.had.co.nz/**

Sr. No.	**Practicals of PGDSP104**
1	1.Develop the R program for Basic Mathematical computation –Square, Square root, exponential etc. 2. Create an object X that stores the value then overwrite the object in by itself divided by Y. Print the result to the console. 3. Create and store a sequence of values from x to y that progresses in steps of 0.3
2	Create and store a three-dimensional array with six layers of a 4 X 2 matrix, filled with a decreasing sequence of values between 4.8 and 0.1 of the appropriate length
3	Extract and store as a new object the fourth- and first-row elements, in that order, of the second column only of all layers of (1).
4	1.Confirm the specific locations of elements equal to 0 in the 10 X 10 identity matrix I10 2.Store this vector of 10 values: foo <- c(7,5,6,1,2,10,8,3,8,2).Then, do the following: i. Extract the elements greater than or equal to 5, storing the result as bar. ii. Display the vector containing those elements from foo that remain after omitting all elements that are greater than or equal to 5.
5	Store the string "Two 6-packs for $12.99". Then do the following: i. Use a check for equality to confirm that the substring beginning with character 5 and ending with character 10 is "6-pack". ii. Make it a better deal by changing the price to $10.99
6	Create a list that contains, in this order, a sequence of 20 evenly spaced numbers between -4 and 4; a 3 X 3 matrix of the logical vector c(F,T,T,T,F,T,T,F,F) filled column-wise; a character vector with the two strings "don" and "quixote"; and a factor vector containing the observations c("LOW","MED","LOW","MED","MED","HIGH"). Then, Extract row elements 2 and 1 of columns 2 and 3, in that order, of the logical matrix.
7	Create and store this data frame as dframe with the fiiels of person,sex,funny in your R workspace.Append the two new records. 3. Write a single line of code that will extract from mydataframe just the names and ages of any records where the individual is female and has a level of funniness equal to Med OR High
8	Create a database with the fields of weight,height and sex then create a plot of weight on the x-axis and height on the y-axis. Use different point characters or colors to distinguish between males and females and provide a matching legend. Label the axes and give the plot a title.
9	Create a plot using ggplot2 for the same database consisting of weight on the x-axis and height on the y-axis. Use different point characters or colors to distinguish between males and females and provide a matching legend. Label the axes and give the plot a title.
10	Write R code that will plot education on the x-axis and income on the y-axis, with both x- and y-axis limits fixed to be [0;100]. Provide appropriate axis labels. For jobs with a prestige value of less than or equal to 80, use a black * as the point character. For jobs with prestige greater than 80, use a blue @.

Class: M.Sc	**Branch**: Data Science	**Semester:** I	
Subject: Data Warehousing & Mining			
Period per Week(Each 60 min)	Lecture	04	
	Practical	04	
Evaluation System		**Hours**	**Marks**
	Semester End Exam	2 hrs.30min	60

	Continuous Internal Assessment	__	40
	Semester End Practical Examination	2 hrs.	50
	Total	__	150

Course: PGDS105	Data Warehousing & Mining **(Credits : 4 Lectures/Week: 4)**	**Lectures**
	Expected Learning Outcomes: After successful completion of this course, students would be able to 1. Explain the operational and decision support system. 2. Evaluate the impact of use and information using knowledge discovery in databases and KDD process models. 3. Summarize the data mining concepts with the help of Apriori algorithm, support, confidence and trees. 4. Construct data models and prototypes needed to gain stakeholder support to achieve business objectives.	
Unit I	**Data Warehouse Fundamentals:** Introduction to Data Warehouse, OLTP Systems, Differences between OLTP Systems and Data Warehouse, Characteristics of Data Warehouse, Components of Data Warehouse, Advantages and Applications of Data Warehouse, Top- Down and Bottom-Up Development Methodology, Tools for Data warehouse development, Data Warehouse Types,	**08L**
Unit-II	**Planning and Requirements:** Introduction: Planning Data Warehouse and Key Issues, Data warehouse Project, Data Warehouse development Life Cycle, The Project Team, Requirements Gathering Approaches: Team organization, Roles, and Responsibilities, Extraction - Transformation - Loading	**10 L**
Unit-III	**OLAP**: Introduction, Characteristics, Advantages, Disadvantages; OLTP vs OLAP, Data cubes, Data cube operations, OLAP types, **Dimensional Modeling**: Dimensional Modeling Basics, E-R Modeling Versus Dimensional Modeling, Data Warehouse Schemas; Star Schema, Inside Dimensional Table, Inside Fact Table, Fact Less Fact Table, Star Schema Keys: Snowflake Schema, Slowly Changing Dimensions	**15 L**
Unit-IV	**Data Mining**: Introduction to Data Mining, The process of knowledge discovery in databases, predictive and descriptive data mining techniques, supervised and unsupervised learning techniques. **Data preprocessing**: Data cleaning, Data transformation, Data reduction, Discretization.	**15 L**
Unit - V	**Classification:** Decision trees, Bayesian classification, **Clustering:** Basic issues in clustering, k-means clustering, Hierarchical clustering- Agglomerative clustering, Divisive clustering, Density-based methods- DBSCAN **Association Rule Mining:** Support, Confidence, Frequent item sets, Apriori algorithm	**12L**
TextBook:		

1. Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals. Paulraj Ponniah
2. Data Mining: Concepts and Techniques, The Morgan Kaufmann Series in Data Management Systems, Han J. and Kamber M. Morgan Kaufmann Publishers, (2000).
3. Data Mining: Introductory and Advanced Topics, Dunham, Margaret H, Prentice Hall (2006)

References:

1. Luis Torgo, Data Mining with R Learning with Case Studies, Second Edition, CRC Press, 2017
2. Building the Data Warehouse, Inmon: Wiley (1993).

Links:

1) **http://www.vssut.ac.in/lecture_notes/lecture1428550844.pdf**
2) **https://lecturenotes.in/subject/32/data-mining-and-data-warehousing-dmdw**

Sr. No.	**Practical of PGDSP105**
1.	Create tables using different applications.
2.	Develop an application to design a warehouse by importing various tables from external sources
3.	a. Develop an application to creating a fact table and measures in a cube b. Develop an application to create dimension tables in a cube and form star schema.
4.	Develop an application to create fact and dimension tables in a cube and form snowflake schema
5.	Develop an application to demonstrate operations like roll-up, drill-down, slice, and dice.
6.	Develop an application to demonstrate processing and browsing data from a cube.
7.	Develop an application to pre-process data imported from external sources.
8.	Pre-process the given data set and hence apply hierarchical algorithms and density based clustering techniques. Interpret the result.
9.	Pre-process the given data set and hence classify the resultant data set using tree classification techniques. Interpret the result.
10.	Create association rules by considering suitable parameters.

Class: M.Sc	**Branch**: Data Science	**Semester:** I	
Subject: Data Structure with Python			
Period per Week(Each 60 min)	Lecture	04	
	Practical	04	
Evaluation System		**Hours**	**Marks**
	Semester End Exam	2 hrs.30min	60
	Continuous Internal Assessment	—	40
	Semester End Practical Examination	2 hrs.	50

	Total	__	150

Course: PGDS106	**Data Structures with Python (Credits : 4 Lectures/Week: 4)**	**Lectures**
	Expected Learning Outcomes: After successful completion of this course, students would be able to 1. Recall the concepts of arrays, strings and algorithms for basic operations. 2. Recognize the concept of stacks, queues, linked list and algorithms for basic operations. 3. Identify the familiarity with major algorithms and data structures 4. Analyze appropriate algorithms and data structures for various applications 5. Formulate the computational complexity of various algorithms	
Unit I	**Abstract Data Types:** Introduction, The Date Abstract Data Type, Bags, Iterators. Application **Arrays:** Array Structure, Python List, Two Dimensional Arrays, Matrix Abstract Data Type, Application **Sets and Maps:** Sets-Set ADT, Selecting Data Structure, List based Implementation, Maps-Map ADT, List Based Implementation, Multi-Dimensional Arrays-Multi-Array ADT, Implementing Multi Arrays, Application **Algorithm Analysis:** Complexity Analysis-Big-O Notation, Evaluating Python Code, Evaluating Python List, Amortized Cost, Evaluating Set ADT, Application **Searching and Sorting:** Searching-Linear Search, Binary Search, Sorting-Bubble, Selection and Insertion Sort, Working with Sorted Lists-Maintaining Sorted List, Maintaining sorted Lists	**15 L**
Unit-II	**Linked lists :** Linear lists, Single Linked List and Chains, Representing Chains, Designing a Chain Class, Chain Manipulation Operations, The Template Class Chain, Implementing Chains with Templates, Chain Iterators ,Chain Operations, Circular List, Doubly Linked Lists, Skip list, Generalized Lists, Representation of Generalized Lists, Recursive Algorithms for Lists, Reference Counts, Shared and Recursive Lists	**15 L**
Unit-III	**Stacks:** Stack ADT, Implementing Stacks-Using Python List, Using Linked List, Stack Applications-Balanced Delimiters, Evaluating Postfix Expressions **Queues:**Queue ADT, Implementing Queue-Using Python List, Circular Array, Using List, Priority Queues- Priority Queue ADT, Bounded and unbounded Priority Queues **Advanced Sorting:** Merge Sort, Quick Sort, Radix Sort, Sorting Linked List	**15 L**
Unit-IV	**Recursion:** Recursive Functions, Properties of Recursion, Its working, Recursive **Hash Table:** Introduction, Hashing-Linear Probing, Clustering, Rehashing, Separate Chaining, Hash Functions **Binary Trees:** Tree Structure, Binary Tree-Properties, Implementation and Traversals, Expression Trees, Heaps and Heapsort,Search Trees, R-Trees & R+ Trees.	**15 L**

TextBook:

1. Data Structure and algorithm Using Python, Rance D. Necaise, 2016 Wiley India Edition
2. Data Structure and Algorithm in Python, Michael T. Goodrich, Robertom Tamassia, M. H. Goldwasser, 2016 Wiley India Edition

References:

1. Data Structure and Algorithmic Thinking with Python-Narasimha Karumanchi, 2015, Careermonk Publications
2. Fundamentals of Python: Data Structures, Kenneth Lambert, Delmar Cengage Learning

Links:

https://lecturenotes.in/subject/81/data-structure-using-c-ds
http://www.cs.yale.edu/homes/aspnes/classes/223/notes.pdf
https://www.smartzworld.com/notes/data-structures-pdf-notes-ds/
https://www.geeksforgeeks.org/data-structures/

Sr. No.	Practicals of PGDSP106
1	Implement Linear Search to find an item in a list.
2	Implement binary search to find an item in an ordered list
3	Implement Sorting Algorithms a. Bubble sort b. Insertion sort c. Quick sort d. Merge sort
4	Implement use of Sets and various operations on Sets.
5	Implement working of Stacks. (pop method to take the last item added off the stack and a push method to add an item to the stack)
6	Implement Program for a. Infix to Postfix conversion b. Postfix Evolution
7	Implement the following a. A queue as a list which you add and delete items from. b. A circular queue. (The beginning items of the queue can be reused).
8	Implement Linked list and demonstrate the functionality to add and delete items in the linked list.
9	Implement Binary Tree and its traversals.
10	Recursive implementation of a. Factorial b. Fibonacci c. Tower of Hanoi

Semester II – Theory

Class: M.Sc	**Branch**: Data Science	**Semester:** II	
Subject: Research in Computing			
Period per Week(Each 60 min)	Lecture	04	
	Practical/ Tutorial	04	
Evaluation System		**Hours**	**Marks**
	Semester End Exam	2 hrs.30min	60
	Continuous Internal Assessment	__	40
	Semester End Practical Examination	2 hrs.	50
	Total	__	150

Course: PGDS201	**Research in Computing (Credits : 4 Lectures/Week: 4)**	**Lectures**
	Expected Course Outcomes After successful completion of this course, students would be able to 1) Develop analytical skills by applying scientific methods. 2) Review the existing research article on Machine learning & Business analytics 3) Survey the specific research areas in field of Computer Science 4) Test & validate the proposed methodology on research problems.	
Unit I	**Introduction:** Role of Business Research, Information Systems and Knowledge Management, Theory Building, Organization ethics and Issues	**12 L**
Unit-II	**Beginning Stages of Research Process:**Problem definition, Qualitative research tools, Secondary data research	**12 L**
Unit-III	**Research Methods and Data Collection:** Survey research, communicating with respondents, Observation methods, Experimental research	**12 L**
Unit-IV	**Measurement Concepts, Sampling and Field work:** Levels of Scale measurement, attitude measurement, questionnaire design, sampling designs and procedures, determination of sample size	**12 L**
Unit-V	**Data Analysis and Presentation:** Editing and Coding, Basic Data Analysis, Univariate Statistical Analysis and Bivariate Statistical analysis and differences between two variables. Multivariate Statistical Analysis.	**12 L**

TextBook:

1. Business Research Methods, William G.Zikmund, B.J Babin, J.C. Carr,Atanu Adhikari, M.Griffin, 8th Edition. 2016.
2. Business Analytics, Albright Winsto, 5th Edition,2015
3.Research Methods for Business Students Fifth Edition, Mark Saunders, 2011.
4.Multivariate Data Analysis, Hair, Pearson, 7th Edition, 2014.

References:

Links:

- http://www.library.auckland.ac.nz/subject-guides/med/pdfs/Hindex%20and%20impact%20factors.pdf
- www.openintro.org/stat/down/OpenIntroStatFirst.pdf

Sr. No.		Practical of PGDSP201
1	A	Write a program for obtaining descriptive statistics of data.
	B	Import data from different data sources (from Excel,csv, mysql, sql server, oracle to R/Python/Excel)
2	A	Design a survey form for a given case study, collect the primary data and analyze it
	B	Perform suitable analysis of given secondary data.
3	A	Perform testing of hypothesis using one sample t-test.
	B	Perform testing of hypothesis using two sample t-test.
	C	Perform testing of hypothesis using paired t-test.
4	A	Perform testing of hypothesis using chi-squared goodness-of-fit test.
	B	Perform testing of hypothesis using chi-squared Test of Independence
5		Perform testing of hypothesis using Z-test.
6	A	Perform testing of hypothesis using one-wayANOVA.
	B	Perform testing of hypothesis using two-wayANOVA.
	C	Perform testing of hypothesis using multivariateANOVA (MANOVA).
7	A	Perform the Random sampling for the given data and analyse it.
	B	Perform the Stratified sampling for the given data and analyse it.
8		Compute different types of correlation.
9	A	Perform linear regression for prediction.
	B	Perform polynomial regression for prediction.
10	A	Perform multiple linear regression.
	B	Perform Logistic regression.

Class: M.Sc	**Branch**: Data Science	**Semester:** II	
Subject: Optimization Techniques			
Period per Week(Each 60 min)	Lecture	04	
	Practical	04	
Evaluation System		**Hours**	**Marks**

	Semester End Exam	2 hrs.30min	60
	Continuous Internal Assessment	__	40
	Semester End Practical Examination	2 hrs.	50
	Total	__	150

Course: PGDS202	**Optimization Techniques (Credits : 4 Lectures/Week: 2)**	**Lectures**
	Expected Learning Outcomes: After successful completion of this course, students would be able to 1) Explain the theory of optimization methods and algorithms. 2) Apply the mathematical results and numerical techniques of optimization theory to concrete data science problems. 3) Apply basic concepts of mathematics to formulate an optimization problem. 4) Analyze and appreciate a variety of performance measures for various optimization problems.	
Unit I	**Introduction to Operations Research** Introduction-Mathematical models of Operation Research-Scope and applications of Operation Research-Phases of Operation Research study-Characteristics of Operation Research-Limitations of Operation Research **Linear Programming** Introduction –Properties of Linear Programming-Basic assumptions-Mathematical formulation of Linear Programming-Limitations or constraints-Methods for the solution of LP Problem-Graphical analysis of LP-Graphical LP Maximization problem-Graphical LP Minimization problem	**15 L**
Unit-II	**Dual Linear Programming** Introduction- Primal and Dual problem -Dual problem properties-Solution techniques of Dual problem-Dual Simplex method-Relations between direct and dual problem-Economic interpretation of Duality	**15 L**
Unit-III	**Transportation and Assignment Models** Introduction: Transportation problem-Balanced-Unbalanced-Methods of basic feasible solutionOptimal solution-MODI method. Assignment problem-Hungarian Method.	**15 L**
Unit-IV	**Network Analysis** Basic concepts-Construction of Network-Rules and precautions-CPM and PERT NetworksObtaining of critical path. Probability and cost consideration. Advantages of Network.	**15 L**
TextBook: 1) Hamdy Taha, Operations Research, 10th edition, Prentice Hall India, 2019. 2) P. K. Gupta and D. S. Hira, Operations Research, S. Chand & co., 2007		
References: 1) S.D. Sharma (2000), Operations Research, Nath & Co., Meerut. Maurice Solient, Arthur Yaspen, Lawrence Fridman, (2003), OR methods and Problems, New Age International Edition. 2) J K Sharma (2007), Operations Research Theory & Applications, 3e, Macmillan India Ltd. P. Sankara Iyer, (2008), Operations Research, Tata McGraw-Hill.		

3) A Ravindran, Don T Philips and James J Solberg, Operations Research: Principles and Practice, 2nd edition, John Wiley and sons, 2007

Links:

https://towardsdatascience.com/tagged/optimization-algorithms
https://www.geeksforgeeks.org/optimization-for-data-science/

Sr. No.	Tutorial of PGDSP206
1	A minimum of 5 problems to be worked out by students in every tutorial class. Another 5 problems per tutorial class to be given as a home work

Class: M.Sc	**Branch**: Data Science	**Semester:** II	
Subject: Statistical Inference			
Period per Week(Each 48 min)	Lecture	04	
	Practical	04	
Evaluation System		**Hours**	**Marks**
	Semester End Exam	2 hrs.30min	60
	Continuous Internal Assessment	—	40
	Semester End Practical Examination	2 hrs.	50

	Total	__	150

Course: PGDS203	**Statistical Inference** **(Credits : 3 Lectures/Week: 3)**	**Lectures**
	Expected Learning Outcomes: After successful completion of this course, students would be able to 1. Recognize several basic types of statistical problems corresponding to various sampling designs. 2. Define null hypothesis, alternative hypothesis, level of significance, test statistic, p value, and statistical significance. 3. Describe the statistical decision-making theory and interpretation. 4. Demonstrate knowledge of the main properties of AR(1),AR(2), ARIMA models. 5. Demonstrate computational skills to implement various statistical inferential approaches.	
Unit I	**Sampling & Sampling Distributions** Introduction to Sampling, Simple random Sampling, Stratified Random Sampling, Cluster Sampling, Concept of Sampling Error, Introduction to Sampling distributions, Student's t distribution, Chi square distribution, Snedecor's F distribution, Interrelations among t, chi-square and F distributions, Central Limit Theorem (Various Versions) and its applications.	**15 L**
Unit II	**Testing of hypothesis** Definitions: population, statistic, parameter, standard error of estimator. Concept of null hypothesis and alternative hypothesis, critical region, level of significance, type I and type II error, one sided and two-sided tests, p- value. Large Sample Tests, Tests based on t, Chi-square and F-distribution. **All tests to be taught using R software. Manual calculations are not expected.**	**15 L**
Unit III	**Analysis of Variance** One Way ANOVA, Two Way ANOVA, Application of ANOVA to test the overall significance of Regression. All topics to be covered using R software. Manual calculations are not expected.	**15 L**
Unit IV	**Time Series** Meaning and Utility. Components of Time Series. Additive and Multiplicative models. Methods of estimating trend: moving average method, least squares method and exponential smoothing method. (single, double and triple), Elimination of trend using additive and multiplicative models. Simple time series models: AR (1), AR (2). Introduction to ARIMA Modelling.	**15 L**

TextBook:

1. Fundamentals of Applied Statistics (3rd Edition), Gupta and Kapoor, S.Chand and Sons, New Delhi, 1987.
2. Time Series Methods, Brockell and Devis, Springer, 2006.

3. Time Series Analysis,4th Edition, Box and Jenkin, Wiley, 2008.

References:

1. Modern Elementary Statistics, Freund J.E., Pearson Publication, 2005.
2. Probability, Statistics, Design of Experiments and Queuing theory with applications Computer Science, Trivedi K.S. ,Prentice Hall of India, New Delhi,2001.
3. Common Statistical Tests, Kulkarni M.B., Ghatpande S.B., Gore S.D., Satyajeet Prakashan,Pune, 1999.
4. Probability And Statistical Inference, 9th Edition, Robert Hogg, Elliot Tanis, Dale Zimmerman, Pearson education Ltd, 2015. A Beginners Guide to R, Alain Zuur, Elena Leno, Erik Meesters, Springer, 2009.
5. Statistics Using R, Sudha Purohit, S.D.Gore, Shailaja Deshmukh, Narosa, Publishing Company.

Links:

1. https://www.youtube.com/watch?v=10cuDKGytMw
2. https://www.tutorialspoint.com/time_series/time_series_moving_average.htm
3. https://otexts.com/fpp2/arima-r.html

Sr. No.	**Practicals of PGDSP203**
1	Write a program on sampling distribution.
2	Write a program on Central Limit Theorem.
3	Write a program on normal test
4	Write a program on t test
5	Write a program on Chi-square
6	Write a program on F-distribution
7	Write a program on One Way ANOVA
8	Write a program on Two Way ANOVA
9	Write a program on AR (1), AR (2)
10	Write a program on ARIMA Modelling

Class: M.Sc	**Branch**: Data Science	**Semester:** II	
Subject: Advanced Python Programming			
Period per Week(Each 60 min)	Lecture	03	
	Practical	02	
Evaluation System		**Hours**	**Marks**
	Semester End Exam	2 hrs.30min	60
	Continuous Internal Assessment	__	40
	Semester End Practical Examination	2 hrs.	50
	Total	__	150

Course: PGDS204	**Advanced Python Programming** **(Credits : 4 Lectures/Week: 2)**	**Lectures**
	Expected Learning Outcomes: After successful completion of this course, students would be able to 1. Explain fundamental understanding of the Python programming language. 2. Describe common Python functionality and features used for data science 3. Illustrate the Object-oriented Programming concepts in Python. 4. Visualize and describe DataFrame structures for cleaning and processing data	
Unit I	**LIST MANIPULATION**:Introduction to Python List· Creating List· Accessing List· Joining List· Replicating List· List Slicing, list comprehension **TUPLES** Introduction to Tuple· Creating Tuples· Accessing Tuples· Joining Tuples· Replicating Tuples· Tuple Slicing· **DICTIONARIES** Introduction to Dictionary· Accessing values in dictionaries· Working with dictionaries· Properties· **Set and Frozeset:** Introduction to Set and Frozenset,Creating Set and Frozenset, Accessing and Joining, Replicating and Slicing **Regular Expressions:** Match function, Search function, Grouping, Matching at Beginning or End, Match Objects , Flags	**15 L**
Unit-II	**Object-Oriented Programming**: Classes and Objects, Creating Classes in Python Creating Objects in Python, The Constructor Method, Classes with Multiple Objects Class Attributes versus Data Attributes, Encapsulation, Inheritance Th Polymorphism. **Functional Programming**: Iterators, Generators, Decorators **Files and Working with Text Data:** Types of Files, Creating and Reading Text Data, File Methods to Read and Write Data, Reading and Writing Binary Files, The Pickle Module, Reading and Writing CSV Files, Python os and os.pathModules, JSON and XML in Python, Processing HTML Files, Processing Texts in Natural Languages	**15 L**

Unit-III	**Working with Tabular Numeric Data(Numpy with Python)**: NumPy Arrays Creation Using *array()* Function, Array Attributes, NumPy Arrays Creation with Initial Placeholder Content, Integer Indexing, Array Indexing, Boolean ArrayIndexing, Slicing and Iterating in Arrays, Basic Arithmetic Operations on NumPy Arrays, Mathematical Functions in NumPy, Changing the Shape of an Array, Stacking and Splitting of Arrays, Broadcasting in Arrays. **Working with Data Series and Frames:** Pandas Data Structures, Reshaping Data, Handling Missing Data, Combining Data, Ordering and Describing Data, Transforming Data, Taming Pandas File I/O **Plotting**: Basic Plotting with PyPlot, Getting to Know Other Plot Types, Mastering Embellishments, Plotting with Pandas	**15 L**

Textbook:

- Michael Urban and Joel Murach, Python Programming, Shroff/Murach, 2016
- Halterman Mark Lutz, Programming Python, O`Reilly, 4th Edition, 2010

References:

1. Wesley J. Chun, "Core Python Programming", Prentice Hall,2006.
2. Mark Lutz, "Learning Python", O'Reilly, 4th Edition, 2009

Links:

https://www.w3schools.com/python
https://docs.python.org/3/tutorial/index.html
https://www.python-course.eu/advanced_topics.php

Sr. No.	Practicals of PGDSP204
1	a. Program with a function that takes two lists L1 and L2 containing integer numbers as parameters. The return value is a single list containing the pairwise sums of the numbers in L1 and L2 b. Program to read the lists of numbers as L1, print the lists in reverse order without using reverse function.
2	Program to find max and min of a given tuple of integers.
3	Write a program that combine lists L1 and L2 into a dictionary.
4	Program to find union, intersection, difference, symmetric difference of given two sets.
5	Write a program for searching, splitting and replacing things based on pattern matching using regular expression.
6	Write programs to parse text files, CSV, HTML, XML and JSON documents and extract relevant data. After retrieving data check any anomalies in the data, missing values etc.
7	Write programs for reading and writing binary files

8	a. Program to implement the inheritance b. Program to implement the polymorphism
9	Write programs to create numpy arrays of different shapes and from different sources, reshape and slice arrays, add array indexes, and apply arithmetic, logic, and aggregation functions to some or all array elements
10	Write programs to use the pandas data structures: Frames and series as storage containers and for a variety of data-wrangling operations, such as: ● Single-level and hierarchical indexing ● Handling missing data ● Arithmetic and Boolean operations on entire columns and tables ● Database-type operations (such as merging and aggregation) ● Plotting individual columns and whole tables ● Reading data from files and writing data to files

Datasets

For this laboratory, appropriate publicly available datasets, can be studied and used. Example:
MNIST (http://yann.lecun.com/exdb/mnist/),
UCI Machine Learning Repository(https://archive.ics.uci.edu/ml/datasets.html),
Kaggle(https://www.kaggle.com/datasets)
Twitter Data

Class: M.Sc	**Branch**: Data Science	**Semester:** II	
Subject: Big Data Analytics			
Period per Week(Each 60 min)	Lecture	04	
	Practical	04	
Evaluation System		**Hours**	**Marks**
	Semester End Exam	2 hrs.30min	60
	Continuous Internal Assessment	__	40
	Semester End Practical Examination	2 hrs.	50
	Total	__	150

Course: PGDS205	**Big Data Analytics** **(Credits : 4 Lectures/Week: 2)**	**Lectures**
	Expected Learning Outcomes: After successful completion of this course, students would be able to 1. Describe the fundamentals of various big data analytics techniques. 2. Design efficient algorithms for mining the data from large volumes. 3. Analyze the HADOOP and Map Reduce technologies associated with big data analytics. 4. Prepare a complete business data analytics solution	
Unit I	**Understanding Big Data:** What is big data , why big data , Data Storage and Analysis, Comparison with Other Systems, Relational Database Management System , Grid Computing, Volunteer Computing, unstructured data, industry examples of big data, web analytics, big data	**15 L**

	and marketing, fraud and big data, risk and big data, big data and healthcare, big data in medicine, advertising and big data, big data technologies, cloud and big data, Crowd sourcing analytics,	
Unit-II	**Big Data MapReduce** MapReduce, Introduction to Map Reduce: The map tasks, Grouping by key, The reduce tasks, Combiners, Details of MapReduce Execution, Word Count MapReduce, Different tools on Big data Platform, Vector data (newspaper article or document search), PageRank Algorithm, Twitter Data Analytic, Social Media mining	**15 L**
Unit-III	**Basics of Hadoop** Data format, introduction to Hadoop, Hadoop ecosystem, analyzing data with Hadoop, scaling out, Hadoop streaming, Hadoop pipes, design of Hadoop distributed file system (HDFS), HDFS concepts, Java interface, data flow, Hadoop I/O, data integrity, compression, serialization, Avro – file-based data structures	**15 L**
Unit-IV	**A General Overview of High-Performance Architecture** – HDFS – MapReduce and YARN – Map Reduce Programming Model, Hive, storage of Hive data (database) in HDFS, Query writing to achieve business tasks, Database management, Query optimization, Views and Partition	**15 L**
Unit-V	**Apache Pig**, What is PIG?, Pig Architecture, Prerequisites, How to Download and Install Pig, Example Pig Script, Data flow programming, Storing data in HDFS / Hood, MongoDB, Database creation, Query building, regular expression	

TextBook:

1. Big Data, Black Book: Covers Hadoop 2, MapReduce, Hive, YARN, Pig, R and Data Visualization, By DT Editorial Services, 2016
2. Programming Hive. By Jason Rutberglen, Dean Wampler, Edward Copriolo, 2012
3. Programming Pig by Anal Gates, 2011
4. MongoDB: The Definitive Guide, by Kristina Chodorow, 2013

References:

1. Hadoop, The Definitive Guide, by Tom White, 2015
2. Mining of Massive Datasets, by Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman, 2015

Links:

http://index-of.co.uk/Big-DataTechnologies/Data%20Science%20and%20Big%20Data%20Analytics.pdf

Sr. No.	Practicals of PGDSP205
1	Write a map-reduce program to count the number of occurrences of each alphabetic character in the given dataset. The count for each letter should be case-insensitive (i.e., include both upper-case and lower-case versions of the letter; Ignore non-alphabetic characters).
2	Write a map-reduce program to count the number of occurrences of each word in the given dataset. (A word is defined as any string of alphabetic characters appearing between non-alphabetic characters like nature's is two words. The count should be case-insensitive. If a word occurs multiple times in a line, all should be counted)
3	Write a map-reduce program to determine the average ratings of movies. The input consists of a series of lines, each containing a movie number, user number, rating and a timestamp.
4	(i)Perform setting up and Installing Hadoop in its two operating modes:

	a. Pseudo distributed, b. Fully distributed. (ii) Use web based tools to monitor your Hadoop setup
5	Implement the following file management tasks in Hadoop: a. Adding files and directories b. Retrieving files c. Deleting files
6	Install and Run Hive then use Hive to create, alter, and drop databases, tables, views, functions, and indexes
7	Install and Run Pig then write Pig Latin scripts to sort, group, join, project, and filter your data.
8	Case Study

Class: M.Sc	**Branch**: Data Science	**Semester:** II	
Subject: Analysis of Algorithms			
Period per Week(Each 48 min)	Lecture	04	
	Practical	04	
Evaluation System		**Hours**	**Marks**
	Semester End Exam	2 hrs.30min	60
	Continuous Internal Assessment	__	40
	Semester End Practical Examination	2 hrs.	50
	Total	__	150

Course: PGDS206	**Analysis of Algorithms** **(Credits : 4 Lectures/Week: 4)**	**Lectures**
	Expected Learning Outcomes: After successful completion of this course, students would be able to 1. Explain the concepts of algorithms for designing good program 2. Implement algorithms using Python 3. Determine how to transform new problems into algorithmic problems with efficient solutions 4. Illustrate algorithm design techniques for solving different problems	
Unit I	**Introduction to algorithm**, Why to analysis algorithm, Running time analysis, How to Compare Algorithms, Rate of Growth, Commonly Used Rates of Growth, Types of Analysis, Asymptotic Notation, Big-O Notation, Omega-Ω Notation, Theta-Θ Notation, Asymptotic Analysis, Properties of Notations, Commonly used Logarithms and Summations, Performance characteristics of algorithms, Master Theorem for Divide and Conquer, Divide and Conquer	**15 L**
Unit II	**Tree algorithms:** What is a Tree? Glossary, Binary Trees, Types of Binary Trees, Properties of Binary Trees, Binary Tree Traversals, Generic Trees (N-ary Trees), Threaded Binary Tree Traversals, Expression Trees, Binary Search Trees (BSTs), Balanced Binary Search Trees, AVL (Adelson-Velskii and Landis) Trees	**15 L**

	Graph Algorithms: Introduction, Glossary, Applications of Graphs, Graph Representation, Graph Traversals, Topological Sort, Shortest Path Algorithms, Minimal Spanning Tree **Selection Algorithms:** What are Selection Algorithms? Selection by Sorting, Partition-based Selection Algorithm, Linear Selection Algorithm - Median of Medians Algorithm, Finding the K Smallest Elements in Sorted Order	
Unit III	**Divide and Conquer** Concept of divide and Conquer, Binary Search (recursive), Quick Sort, Merge sort **Greedy Method** Fractional Knapsack problem, Optimal Storage on Tapes, Huffman codes, Concept of Minimum Cost Spanning Tree, Prim's and Kruskal's Algorithm **Dynamic Programming** The General Method, Principle of Optimality, Matrix Chain Multiplication, 0/1 Knapsack Problem, Concept of Shortest Path, Single Source shortest path, Dijkstra's Algorithm, Bellman Ford Algorithm, Floyd-Warshall Algorithm, Travelling Salesperson Problem	**15 L**
Unit IV	**Branch & Bound** Introduction, Definitions of LCBB Search, Bounding Function, Ranking Function, FIFO BB Search, Traveling Salesman problem Using Variable tuple. **Decrease and conquer** Definition of Graph Representation, BFS, DFS, Topological Sort/Order, Strongly Connected Components, Biconnected Component, Articulation Point and Bridge edge **Problem Classification** Basic Concepts: Deterministic Algorithm and Non deterministic, Definitions of P, NP, NP-Hard, NP-Complete problems, Cook's Theorem (Only Statement and Significance)	**15 L**

TextBook:

1. Data Structure and Algorithmic Thinking with Python, Narasimha Karumanchi , CareerMonk Publications, 2016
2. Introduction to Algorithm, Thomas H Cormen, PHI

Additional References:

1. Data Structures and Algorithms in Python, Michael T. Goodrich, Roberto Tamassia, Michael H. Goldwasser, 2016, Wiley
2. Fundamentals of Computer Algorithms, Sartaj Sahni and Sanguthevar Rajasekaran Ellis Horowitz, Universities Press

Links:

1. **https://www.tutorialspoint.com/data_structures_algorithms/**
2. **https://www.javatpoint.com/data-structure-tutorial**

Sr. No.	**Practicals of PGDSP202**
1	Write a Python program to perform matrix multiplication. Discuss the complexity of the algorithm used.
2	Write a Python program to sort n names using Quick sort algorithm. Discuss the complexity of the algorithm used.
3	Write a Python program to sort n numbers using Merge sort algorithm. Discuss the complexity of algorithm used
4	Write a Python program for inserting an element into a binary tree.
5	Write a Python program for deleting an element (assuming data is given) from a binary tree.
6	Write a Python program for checking whether a given graph G has a simple path from source s to destination d. Assume the graph G is represented using adjacency matrix

7	Write a Python program for finding the smallest and largest elements in an array A of size n using the Selection algorithm. Discuss Time complexity
8	Write a Python program for finding the second largest element in an array A of size n using Tournament Method. Discuss Time complexity.
9	Write a Python program for implementing Huffman Coding Algorithms. Discuss the complexity of algorithm
10	Write a Python program for implementing Strassen's Matrix multiplication using Divide and Conquer method. Discuss the complexity of the algorithm.