Department of Statistics
Institute of Social Sciences

M.Sc. (Data Science)	
Program Educational Objectives (PEO's)	The primary objective of the MSc. program in Data Science is to develop skilled professional workforce that is prepared to address the increasing needs in the rapidly expanding area of big data analytics. The program aims to provide skills in quantitative data analyses, data mining, data modeling and prediction, data storage and management, big data processing, data visualization, multimedia big data, programming and communication skills. Software based courses/ training and a large number of practical case studies have been integrated in the program to boost the learner confidence and market acceptability.
Program Outcomes (PO's) :	On completion of M.Sc. Data Science programme, graduates will be able to • Become a skilled Data Scientist in industry, academia, or government organizations. • Use specialist software tools for data storage, analysis and visualization. • Independently carry out research/investigation to solve practical problems

Syllabus

M.Sc. (Data Science): I Semester		
Course Code: MDS 101	**Descriptive Statistics for Data Science**	**Credit: 4** **Max Marks: 60+40 =100**
Course objectives:	To have basic idea about the presentation and analysis of the data.	
Course outcomes:	On successful completion of this course, the students will be able to • Present data in different graphical forms • Formulate the data and draw inference using parametric and non-parametric tests.	
Unit I	Data types: discrete, continuous, univariate, multivariate, binary, character, factor. Exploratory data analysis: Box plot, Stem and leaf, dot plot, central tendency measures, measures of dispersion	
Unit II	Measures of association: consistency and independence of data with special reference to attributes. Karl Pearsons's correlation coefficient: simple, partial and multiple correlations. Regression analysis, covariance matrix, correlation matrix	
Unit III	Formulation of statistical hypothesis. Hypothesis testing Student: t-test: single mean test, independent sample t-test, paired sample t-test. Chi-square test: goodness of fit, test of independence, test for correlation and test for dispersion. Confidence limits and confidenceintervals.	
Unit IV	ANOVA: Introduction of linear models, one way and two way ANOVA. Non-parametric tests: Sign test, signed rank test, Kolmogrov-Smirnov test, Test of independence (run test). Two sample problem: Wilcoxon Mann-Whitney test, Median test, Kolmogrov-Smirnov test, run test.	
	Books Recommended: 1. Lander, J.P. (2017): *R for Everyone-Advanced Analytics and Graphics.* Pearson Education.	

	2. Dalgaard, P. (2008): *Introductory Statistics with R.* Springer. 3. Robinson, A.P. and Hamann, J. D. (2010): *Forest Analytics with R-An Introduction.* Springer. 4. Wolfe, Douglas A. and Schneider, G. (2017): *Intuitive Introductory Statistics.*Springer. 5. Mood, A.M, Graybill F.A and Boes D.C. (2008): *Introduction to the Theory of Statistics*. McGraw Hill, New Delhi. 6. Montgomery D.C and Runger G.C. (2013): *Applied Statistics and Probability for Engineers.* Wiley India, New Delhi.

M.Sc. (Data Science): I Semester		
Course Code: MDS 102	**Linear Algebra and Matrix Computation**	**Credit: 4** **Max Marks: 60+40 =100**
Course objectives:	To study the theory of linear algebra in the light of data science.	
Course outcomes:	On successful completion of this course, the students will be able to • Conceptualized the fundamentals of vectors and metrics in linear algebra. • Apply these concepts in the field of data science.	
Unit I	Vectors and Matrices: Operations on matrices and vectors, special matrices, partitioned matrices, linear and quadratic forms, rank of matrices, determinant of a matrix, inverse matrices and applications in linear models and multivariate methods.	
Unit II	Vector and Matrix Calculus: Differentiation of scalar with respect to a vector, differentiation of a scalar with respect to a matrix, differentiation of a vector with respect to vector, differentiation of a matrix with respect to scalar, eigen analysis of real symmetric matrices, properties of eigen values, statistical application to principal component analysis, use of eigen analysis in constrained optimization.	
Unit III	Matrix decomposition: QR, LU, LDU, Cholesky decomposition, generalized inverse, solution of linear equations, Hadamard products, Kronecker products and vector operator.	
Unit IV	Applications: Applications to multivariate normal distribution, estimation of its parameters and testing of hypothesis for mean vector and variance covariance matrix, principal componentanalysis, linear discriminant analysis, linear models.	
	Books Recommended: 1. Fieller, N. (2015): *Basics of Matrix Algebra for Statistics with R.* CRC Press. *2.* Gentle, J. E. (2017): *Matrix Algebra- Theory, Computations and Applications in Statistics*. Second edition. Springer. 3. Boyd, S. and Vandenberghe, L. (2018): *Introduction to Applied Linear Algebra Vectors*. Matrices and Least Squares, Cambridge University Press. 4. Allaire, G. and Kaber, S. M. (2008): *Numerical Linear Algebra, Texts in Applied Mathematics.* Springer. 5. Hogben, L. (2014). *Handbook of Linear Algebra.* CRC Press/Taylor & Francis Group. 6. Friedberg, S., Insel, A., and Spence, L. (2019).*Linear Algebra.* 5/e, Pearson.	

M.Sc. (Data Science): I Semester		
Course Code: MDS 103	**Regression Analysis and Predictive Modeling**	**Credit: 4 Max Marks: 60+40 =100**
Course objectives:	To introduce the advanced regression analysis and to study theprediction based modeling.	
Course outcomes:	On successful completion of this course, the students will be able to • Formulate the linear models in the field of data science. • Use these models in real life problems for prediction.	
Unit I	Linear Models: Simple linear regression, estimating the coefficients, accuracy of the coefficient estimates, model accuracy. Multiple linear regression: regression coefficients, qualitative Predictors, comparison of linear regression with K-Nearest Neighbours	
Unit II	Generalized linear models: logistic regression, estimating the regression coefficients, multiple logistic regression, logistic regression for >2 response classes, Poisson regression, other generalised linear models.	
Unit III	Model diagnostics: residuals, comparing models, cross validations: validation set approach, leave-one-out cross-validation, k-fold cross-validation. Bootstrap: validation set approach, leave-one- out, k-fold cross-validation, stepwise variable selection.	
Unit IV	Regularization: subset selection, shrinkage methods, reduction methods, Elastic net, Bayesian shrinkage, non-linear least squares, model selection.	
	Books Recommended: 1. Lander, J.P. (2017). *R for Everyone-Advanced Analytics and Graphics*. Pearson Education. 2. James, G., Witten, D., Hastie, T. and R. Tibshirani (2013). *An Introduction to Statistical Learning: with Applications in R.* Springer. 3. Montgomery, D.C., Peck, E.A. and Vining, G.G. (2016). *Introduction to Linear Regression Analysis.* Third Ed., Wiley India Pvt. Ltd.. 4. Draper, N.R. and Smith, H. (2015):*Applied Regression Analysis.*WILEY India Pvt. Ltd. New Delhi; Third Edition, 2015.	

M.Sc. (Data Science): I Semester		
Course Code: MDS 104	**Probability and Distribution Theory**	**Credit: 4 Max Marks: 60+40 =100**
Course objectives:	To introduce the concepts of probability and distribution theoryusing R	
Course outcomes:	On successful completion of this course, the students will be able to • Apply the results of probability in data science using R. • Understand the concepts of Probability and its distribution using R.	
Unit I	Empirical definition of Probability, random variable: discrete and continuous. Discrete probability distributions: Uniform, Bernoulli, Binomial, Poisson, Geometric, Negative Binomial. Simulations from distributions in R.	
Unit II	Continuous probability distributions: Uniform, normal, exponential, gamma, weibull, lognormal. Simulations from distributions in R.	
Unit III	Sequence of random variables, laws of convergence in probability, Chebychev's inequality,SLLN,	

	WLLN, Central limit theorem, Illustration using the replicate function of R
Unit IV	Sampling distributions: statistic, standard test statistic: t, chi-square, F. Interrelationships betweent, chi-square and F. Density estimation. Illustrations of sampling distributions using simulation tools.
	Books Recommended: 1. Hogg, R., Tanis, E. and Zimmerman, D. (2019): *Probability and Statistical Inference.* Pearson Education India. 2. Lander, J.P. (2017):*R for Everyone-Advanced Analytics and Graphics*. Pearson Education. 3. Murdoch, D. and Braun, J. (2020): *A First Course in Statistical Programming with R*. Cambridge University Press. 4. Miller, I. and Miller, M. (2017): *John E. Freund's; Mathematical Statistics*. Pearson. 5. Shanmugam, R. and Chattamvelli, R. (2015): *Statistics for scientists and engineers.* John Wiley

M.Sc. (Data Science): I Semester		
Course Code: MDS 105	**Lab Work Based on MDS 101, 102, 103, 104**	**Credit: 4** **Max Marks: 60+40 =100**
Course objectives:	To learn real life/industry applications of theory	
Course outcomes:		

M.Sc. (Data Science): II Semester		
Course Code: MDS 201	Programming for Data Science with R	**Credit: 4** **Max Marks: 60+40 =100**
Course objectives:	To Introduce the elementary and advanced concepts of R language.	
Course outcomes:	On successful completion of this course, the students will be able to • Describe statistical modeling using R • Apply these modeling tools in Statistical/Machine learning	
Unit I	Data exploration and data wrangling: visualization with *ggplot2*, data transformation with *dplyr,* Exploratory Data Analysis, R studio projects, data wrangling, creating tibbles, data import with readr, tidy data with tidyr, relational data with dplyr, strings with stringr, factors with forcats, dates and times with lubridate	
Unit II	Programming: pipes with magrittr, functions, function arguments, return values, vectors, atomic vectors, recursive vectors, augmented vectors. Iteration with purr, for loops versus functionals, the map functions, other patterns of for loops.	
Unit III	Modeling: model basics with modelr, visualizing models, formulas and model families, model building, many models with purrr and broom, gapminder, creating list-columns, making tidy data with broom, making tidy data with broom.	
Unit IV	Communication: R markdown, basics, code chunks, YAML header, graphics for communication with *gglpot2*, output options, documents, notebooks, presentations, dashboards, websites.	
	Books Recommended: 1. Wickham, H. and Grolemund, G. (2016): *R for Data Science Import, Tidy, Transform,*	

	Visualize, and Model Data. O'Reilly. 2. Lander, J.P. (2017): *R for Everyone-Advanced Analytics and Graphics*. Pearson Education. 3. Wickham, H. (2014): *Advanced R.* CRC Press. Gandrud, C. (2020): *Reproducible Research with R and R Studio.* 3rd edition, CRC Press.

M.Sc. (Data Science): II Semester		
Course Code: MDS 202	Fundamental of Data Base Management System	**Credit: 4** **Max Marks: 60+40 =100**
Course objectives:	To Provide Knowledge of data base management through R.	
Course outcomes:	On successful completion of this course, the students will be able to • Know the objectives of data management • To extract data from various data bases	
Unit I	RDBAM and SQL RDBMS overview, queries on one table, joins, self joins, inner-join, outer- join, multiple relations between tables, set operations, aggregate operations, efficient queries, structured query language	
Unit II	Commands in SQL, data types in SQL, data manipulation and data processing with SQL	
Unit III	Graphical Practice: graphical excellence, graphical integrity, theory of data graphics: data-ink and graphical redesign, data-ink maximization and graphical design, multifunctioning graphical elements, data density and small multiples, aesthetics and technique in data graphical design.	
Unit IV	Import/export data from various data bases in/from R, understanding features/installation of MYSQL, SQLite, excel, MongoDB, postgre SQL, hive, Hbase, installing R interfaces of popular data bases.	
	Books Recommended: 1. Walkowiak, S. (2016): B*ig Data Analytics with R.* Packt open source. 2. Sumathi, S. and Esakkirajan, S. (2007): *Fundamentals of Relational Database Management Systems.*Springer. 3. Edward R. Tufte (2001):*The Visual Display of Quantitative Information.* Prajapati, V. (2013). *Big Data Analytics with R and Hadoop.* Packt open source	

M.Sc. (Data Science): II Semester		
Course Code: MDS 203	**Bayesian Data Analysis**	**Credit: 4** **Max Marks: 60+40 =100**
Course objectives:	To introduce the elementary and advanced concepts of Bayesian Data Analysis.	
Course outcomes:	On successful completion of this course, the students will be able to • Analyse the data through the techniques of Bayesian inference • Apply the Bayesian inference to real life scenario.	
Unit I	Fundamentals of Bayesian Inference: probability and inference. single-parameter models, introduction to multi-parameter models, asymptotic and connections to non-Bayesian approaches, hierarchical models.	
Unit II	Fundamentals of Bayesian data analysis, model checking, evaluating, comparing, and expanding models, posterior predictive checking, graphical posterior predictive checks, evaluating,comparing, and expanding models, information criteria and cross-validation, model comparison based on predictive performance, model expansion.	
Unit III	Bayesian computation: Introduction to Bayesian computation, Markov chain Monte Carlo simulations, Gibbs sampler, Metropolis and Metropolis-Hastings algorithms, convergence ofchains, Effective number of simulation draws, Hamiltonian Monte Carlo, Hamiltonian dynamics for a	

	simple hierarchical model with STAN.
Unit IV	Regression Models: Bayesian approach to regression models with single multiple predictors. Hierarchical linear models with Bayesian approach and generalized linear models.
	Books Recommended: 1. Gelman, A., John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari and Donald B. Rubin (2014): *Bayesian Data Analysis*. 3rd Edition. Chapman and Hill. 2. Peter, D. H. (2009): *A First Course in Bayesian Statistical Methods.* Springer-Verlag. NewYork. 3. McElreath, R. (2019): *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. (Chapman & Hall/CRC Texts in Statistical Science). 4. Kruschke, J. (2014):*Doing Bayesian Data Analysis: A Tutorial with R, JAGS and Stan*. Academic Press. 5. Rasmussen, C.E. and Williams, C.K.I. (2006):*Gaussian Processes for Machine Learning*. MIT Press (2006).

M.Sc. (Data Science): II Semester		
Course Code: MDS 204	**Machine Learning**	**Credit: 4** **Max Marks: 60+40 =100**
Course objectives:	To introduce the basis concepts of machine learning	
Course outcomes:	On successful completion of this course, the students will be able to • Describe the concepts of machine learning • Apply the machine learning tools in data science.	
Unit I	Machine learning: Concept and issues, Supervised versus unsupervised learning, Regressionversus Classification problem, Algorithms versus Models, Model training: regression and classification models, , model assessment , bias-variance trade-off, hyper parameter tuning, cross validation, ROC curves.	
Unit II	Tree based methods: Basics of decision trees, a simple tree, tree entropy and information gain, Trees versus linear models, pros and cons of trees, overfitting, pruning a tree, Trees versus linear models, bagging, random forests, boosting, fitting of classification and regression trees.	
Unit III	Support vector machines (SVMs): Overview, separating hyperplane, maximal margin classifier, support vector classifier (SVC): linear classification and classification with non-linear decision boundaries, SVM versus SVC, SVM with more than 2 classes: One-versus-One and One-versus- All case, kernel functions.	
Unit IV	Neural Networks: Overview, single and multilayer neural networks, neural networks for regression and classification. kNN classifier and k means clustering as machine learning tools.	
	Books Recommended: 1. Lewis, N.D. (2017): *Machine Learning Made Easy with R: An Intuitive Step by Step Blueprint for Beginners.* CreateSpace Independent Publishing Platform. 2. Burger, S.V. (2018): *Introduction to Machine Learning with R: Rigorous mathematical modeling*. O Reilly. Lantz, B. (2019): *Machine Learning with R: Expert Techniques for Predictive Modeling*.Packt Publications, 3rd edition.	

M.Sc. (Data Science): II Semester		
Course Code: MDS 205	**Lab Work Based on MDS 201, 202, 203, 204**	**Credit: 4** **Max Marks: 60+40 =100**
Course objectives:	To learn real life/industry applications of theory	

Course outcomes:	

M.Sc. (Data Science): III Semester		
Course Code: MDS 301	**Multivariate Methods in Data Science**	**Credit: 4** **Max Marks: 60+40 =100**
Course objectives:	To introduce the elementary and advanced concepts of multivariateanalysis.	
Course outcomes:	On successful completion of this course, the students will be able to • Describe the methods and techniques of multivariate analysis in data science. • Apply these methods & techniques in real life problem.	
Unit I	Multivariate data and its summaries: numeric and graphical, outlier detection and inferences, multivariate distance measures, reducing data complexity, Principal component analysis (PCA), exploratory factor analysis, multidimensional scaling.	
Unit II	Discriminant Analysis: discrimination using distance measures, normality assessment, transformations, discrimination using Bayes theorem, linear discriminant analysis, quadratic discriminant analysis, stepwise discriminant function analysis, canonical discriminant function, logistics regression.	
Unit III	Confirmatory factor analysis and structured equation modeling: motivation for structured models, scale assessment, structured equation models, partial least squares, SEM with large samples.	
Unit IV	Segmentation methods, Clustering: types, hierarchical methods, limitations; PCA with Cluster analysis, Classification methods.	
	Books Recommended: 1. Bryan F.J. Manly, Jorge A. Navarro Alberto (2017): *Multivariate Statistical Methods: A Primer*. Chapman and Hall/CRC. Fourth edition. 2. Chapman, C. and McDonnell Feit, E. (2019): *R for Marketing Research and Analytics*. II Edition. Springer. 3. Hardly W.K. and Simor L. (2015): *Applied Multivariate Statistical Analysis,* 4th Edition, Springer-Verlag. 4. Richard A. Johnson and Dean W. Wichern (2019):*Applied Multivariate Statistical Analysis*. Prentice hall India, 7th Edition.	

M.Sc. (Data Science): III Semester		
Course Code: MDS 302	**Programming for Data Science with Python**	**Credit: 4** **Max Marks: 60+40 =100**
Course objectives:	To introduce the basic and advanced elements of Pythonprogramming	
Course outcomes:	On successful completion of this course, the students will be able to • Demonstrate the programming skills in Python • Apply the Python programming for data analysis	
Unit I	Python Language Basics, IPython, and Jupyter Notebooks: IPython Basics, Tab Completion, Introspection, Matplotlib integration, executing code from the clipboard. python language basics, scalar types. data structures and sequences, tuple, list, set, and dict comprehensions, namespaces, scope, and local functions, returning multiple values, currying: partial argument application, generators, errors and exception handling, errors handling. files and the operating system.	
Unit II	The NumPy ndarray: a multidimensional array object, creating ndarrays, data types for ndarrays, arithmetic with numpy arrays, basic indexing and slicing, transposing arrays and swapping axes, universal functions: fast element-wise array functions, array-oriented programming with arrays, file	

	input and output with arrays, pseudorandom number generation. introduction to pandas data structures, series, data frame, reindexing, indexing, selection and filtering, integer indexes, sorting and ranking. data loading, storage, and file formats reading and writing data in text format, JSON Data, XML and HTML: Web Scraping, Binary Data Formats.
Unit III	Data Cleaning and preparation: handling missing data, data transformation, string manipulation. data wrangling: hierarchical indexing, combining and merging datasets, reshaping and pivoting. plotting and visualization: matplotlib api primer, plotting with pandas and seaborn, other python visualization tools. data aggregation and group operations: groupby mechanics, data aggregation, apply: general split-apply-combine, pivot tables and cross-tabulation.
Unit IV	Time series: date and time data types and tools, time series basics, date ranges, frequencies, and shifting, time zone handling, periods and period arithmetic, resampling and frequency conversion. advanced pandas: categorical data, advanced groupby use, techniques for method chaining. introduction to modeling libraries in python: interfacing between pandas and model code, creating model descriptions with patsy, introduction to statsmodels, scikit-learn.
	Books Recommended: 1. McKinney, W. (2018): *Python for Data Analysis.* O'Reilly. 2. Nelli, F.(2018): *Python Data Analytics: with Pandas, NumPy and Matplotlib*. A press. 3. David J. Pine (2019): *Introduction to Python for Science and Engineering.* CRC Press. 4. Jake vanderPlas (2017): *Python Data Science Handbook – Essential Tools for Working with Data*.O'Really Media. Johansson, R.(2019):*Numerical Python-Scientific Computing and Data Science Applications withNumPy, SciPy and Matplotlib.* A press,

M.Sc. (Data Science): III Semester		
Course Code: MDS 303	**Operations Research**	**Credit: 4** **Max Marks: 60+40 =100**
Course objectives:	To introduce the basic and advanced concept of Operations Research	
Course outcomes:	On successful completion of this course, the students will be able to • Formulate the real life decision making problem into mathematical model. • Solve complex decision making problems through various techniques ofOperations Research.	
Unit I	Introduction to Operation Research-Scope, Applications and Limitations of Operation Research, Linear Programming Problem (LPP), Properties of Linear Programming, basic assumptions, Mathematical formulation of LPP, Solution of LPP-Simplex method (Big –M method, Two phase method), Revised simplex method	
Unit II	Primal and dual problem, dual simplex method, economic interpretation of duality, sensitivity analysis, parametric programming	
Unit III	Transportation problem (Balanced/Unbalanced), Methods of basic feasible solution, MODI method, Assignment problem, Hungarian Method. Basic concepts, Construction, Rules, precautions and advantages of network, CPM and PERT Networks, Obtaining of critical path, Introduction to game theory, Two Person Zero-Sum game-Solution of games with saddle points and without saddle points, games, Dominance principle	
	Unit IV Introduction to Integer Programming Problem (IPP), Formulation and Applications of IPP, Cutting Plane methods: Fractional cut, Mixed Integer Methods, Branch and Bound method: Dakin's approach, Multi-objective Optimization, Goal programming, Fuzzy goal programming	
	Books Recommended:	

	1. Taha, H. (2019): *Operations Research.* 10th edition, Prentice Hall India. 2. Gupta, K. P. and Hira, D.S. (2007): *Operations Research.* S. Chand & co.. 3. Salkin, H.M. (1975): *Integer Programming.* Addison Wesky. 4. Rao, S.S. (1989): *Optimization: Theory and Applications.* Wiley Eastern. 5. Hadley G. (1970): *Nonlinear and Dynamic Programming*. Addison Wesley. 6. Bazara and Shetty (1979): *Nonlinear Programming*. John Wiley. **Hillier and Lieberman (1991): *Introduction Mathematical Programming*. McGraw Hill.**

M.Sc. (Data Science): III Semester		
Course Code: MDS 304	**Time Series Analysis**	**Credit: 4** **Max Marks: 60+40 =100**
Course objectives:	To provide knowledge of elementary and advanced concepts of Time Series Analysis	
Course outcomes:	On successful completion of this course, the students will be able to • Demonstrate the concepts of time series analysis • Forecasts with valid conclusions based on appropriate time series data.	
Unit I	Characteristics of Time Series: the nature of time series data, time series statistical models, measures of dependence, stationary time series, estimation of correlation, vector-valued and multidimensional series. time series regression and exploratory data analysis: classical regression in the time series context, exploratory data analysis, smoothing in the time series context.	
Unit II	ARIMA Models: autoregressive moving average models, difference equations, autocorrelation and partial autocorrelation, forecasting, estimation, integrated models for nonstationary data, building ARIMA Models, regression with autocorrelated errors, multiplicative seasonal ARIMA Models.	
Unit III	Spectral Analysis and Filtering: cyclical behavior and periodicity, the spectral density, periodogram and discrete fourier transform, nonparametric spectral estimation, parametric spectral estimation, multiple series and cross-spectra, linear filters, GARCH Models, Long Memory ARMA and Fractional Differencing, Unit Root Testing.	
Unit IV	State Space Models: Linear Gaussian Model, filtering, smoothing, and forecasting, maximum likelihood estimation, missing data modifications, structural models: signal extraction and forecasting, state-space models with correlated errors, bootstrapping state space models.	
	Books Recommended: 1. Shumway, R.H. and Stoffer, D.S. (2017): *Time Series Analysis and Its Applications: With R Examples*. Fourth Edition.Springer. 2. Montgomery, D.C., Jennings, C. and Kulahci, M. (2016): *Introduction to Time Series Analysis and Forecasting*. Second Ed., Wiley. Box, G., Jenkins, G.M. Reinsel, G.C. and Ljung, G. (2016): *Time Series Analysis: Forecastingand Control*. Fifth Ed., Wiley.	

M.Sc. (Data Science): III Semester		
Course Code: MDS 305	**Lab Work Based on MDS 301, 302, 303, 304**	**Credit: 4** **Max Marks: 60+40 =100**
Course objectives:	To learn real life/industry applications of theory	
Course outcomes:		

M.Sc. (Data Science): IV Semester			
Course Code: MDS 401		**Big Data Analytics**	**Credit: 4** **Max Marks: 60+40 =100**
Course objectives:		To study the specialized aspects of big data analytics.	
Course outcomes:		On successful completion of this course, the students will be able to • Identify big data and its real life implications • Analyze the problem of big data with the help of R and Hadoop	
Unit I	Introduction big data, big bata toolbox: Hadoop and Spark, traditional limitations with the softwares, expanding memory with the bigmemory package, parallel computing, Boosting R performance for big data.		
Unit-II	Hadoop architecture: Hadoop Distributed File System, MapReduce framework, single-node Hadoop in Cloud, HDInsight-a multi-node Hadoop cluster on Azure, smart energy meter readingsanalysis example-using R on HDInsight cluster		
Unit-III	Relational Database Management Systems: SQLite with R, MariaDB with R on amazon EC2 instance, PostgreSQL with R on amazon RDS. Non-Relational (NoSQL) Databases: processing big data using MongoDB, HBase.		
Unit-IV	Spark for big data analytics: Spark with a multi-node HDInsight cluster, Reading the data into HDFS and Hive. Machine learning methods for big data: GLM examples, naive Bayes with H2O on Hadoop, neural networks with H2O on Hadoop.		
	Books Recommended: 1. Walkowiak, S. (2016): B*ig Data Analytics with R.* Packt open source. 2. Pries, K. H. and Dunnigan, R. (2015): *BIG DATA ANALYTICS- A Practical Guide for Managers*. CRC Press. Prajapati, V. (2013): *Big Data Analytics with R and Hadoop.* Packt open source.		

M.Sc. (Data Science): IV Semester			
Course Code: MDS 402		**Marketing Research and Analytics**	**Credit: 4** **Max Marks: 60+40 =100**
Course objectives:		To improve the quality of decision making of marketing research through the study of relevant data and information	
Course outcomes:		On successful completion of this course, the students will be able to • Evaluate and design marketing research problems. • Analyze the marketing research models with the help of R Programming.	
Unit I	Reducing Data Complexity: consumer brand rating data, principal component analysis and perceptual maps, exploratory factor analysis, multidimensional scaling. Linear modeling: handling highly correlated variables, linear models for binary outcomes: logistic regression, hierarchical models, Bayesian hierarchical linear models, quick comparison of the effects.		
Unit-II	Confirmatory factor analysis and structural equation modeling: the motivation for structural models, scale assessment: confirmatory factor analysis (cfa), general models: structural equation models, the partial least squares (pls) alternative.		
Unit-III	Segmentation: clustering and classification: segmentation philosophy, segmentation data,clustering, classification, prediction: identifying potential customers. association rules for market basket analysis: the basics of association rules, retail transaction data: market baskets, finding and visualizing association rules, rules in non-transactional data: exploring segments again.		
Unit-IV	Choice modeling: choice-based conjoint analysis surveys, simulating choice data, fitting a choice model, adding consumer heterogeneity to choice models, hierarchical bayes choice models,design of choice-based conjoint surveys. behavior sequences: web log data, basic event statistics,		

	identifying sequences (sessions), markov chains for behavior transitions.
	Books Recommended: 1. Chapman, C. and McDonnell Feit, E. (2019): *R for Marketing Research and Analytics*. II Edition. Springer. 2. Ohri, A, (2013): *R for Business Analytics.* Springer. Malhotra, N.K., Daniel, N. and Birks, D.F. (2017*)*: *Marketing Research: An Applied Approach.* Pearson Education Limited.

M.Sc. (Data Science): IV Semester		
Course Code: MDS 403	**Spatial Statistics for Remotely Sensed Images**	**Credit: 4** **Max Marks: 60+40 =100**
Course objectives:	To study data analysis with reference to spatial data.	
Course outcomes:	On successful completion of this course, the students will be able to • Understand spatial data, image, image format, types of images • Classifications and analysis of supervised and unsupervised images.	
Unit I	Spatial data: An overview, some measurements on spatial data, Image, Image formats, types of images, basic image manipulation, reading, saving and displaying an image, image correction: radiometric and geometric, image registration.	
Unit II	Image enhancement: Point wise intensity transformation, histogram equalization and matching, linear smoothing, non-linear smoothing, low-pass filtering, high-pass filtering.	
Unit III	Image transformations: Spectral indices, Principal component transformation, Fourier transformation, wavelet transformation, image restoration, Image segmentation, Edge/Regionbased segmentation, spectral clustering for image segmentation, edge detection	
Unit IV	Supervised and Unsupervised image classification: Maximum likelihood classifier, Distance based classification, parallelepiped classifier, kNN classifier, linear discrimination, support vector classifiers, neural network classifiers, tree based classifiers, clustering based unsupervised classification, k-means clustering, fuzzy k means clustering, agglomerative hierarchical clustering, Gaussian mixture clustering, Kohonen self-organizing map, hybrid supervised/ unsupervised classification, classification accuracy assessment.	
	Books Recommended: 1. Canty, M.J. (2019): *Image Analysis, Classification and Change Detection in Remote Sensing-With Algorithms for Python*. 4th edition, CRC Press. 2. Richards, J.A., Jia, X (2006): *Remote Sensing Digital Image Analysis: An introduction*. 4th edition, Springer. Petrou, M., Petrou, C.(2010): *Image Processing: The fundamentals*. 2nd edition, Wiley	

M.Sc. (Data Science): IV Semester		
Course Code: MDS 404	**Cloud Computing**	**Credit: 4** **Max Marks: 60+40 =100**

Course objectives:	To provide the conceptual knowledge of Cloud Computing
Course outcomes:	On successful completion of this course, the students will be able to • Develop technological foundation of cloud computing • Make innovations using cloud computing
Unit I	Systems Modeling, Clustering and Virtualization: Distributed System Models and Enabling Technologies, Computer Clusters for Scalable Parallel Computing, Virtual Machines and Virtualization of Clusters and Data centers. Foundations: Introduction to Cloud Computing, Migrating into a Cloud, Enriching the „Integration as a Service" Paradigm for the Cloud Era, The Enterprise Cloud Computing Paradigm..
Unit II	Infrastructure as a Service (IAAS) & Platform and Software as a Service (PAAS / SAAS):Virtual machines provisioning and Migration services, On the Management of Virtual machines for Cloud Infrastructures, Enhancing Cloud Computing Environments using a cluster as a Service, Secure Distributed Data Storage in Cloud Computing. Aneka, Comet Cloud, T-Systems, Workflow Engine for Clouds, Understanding Scientific Applications for Cloud Environments
Unit III	Monitoring, Management and Applications: An Architecture for Federated Cloud Computing, SLA Management in Cloud Computing, Performance Prediction for HPC on Clouds, Best Practices in Architecting Cloud Applications in the AWS cloud, Building Content Delivery networks using Clouds, Resource Cloud Mashups.
Unit IV	Governance and Case Studies: Organizational Readiness and Change management in the Cloud age, Data Security in the Cloud, Legal Issues in Cloud computing, Achieving Production Readiness for Cloud Services.
	Books Recommended: 1. Vecciola, B. and Selvi (2017): *Mastering Cloud Computing: Foundations and Applications Programming*. Tata McGraw Hill. 2. Rittinghouse and Ransome (2009): *Cloud Computing: Implementation, Management and Security.* CRC Press, 3. Doss, A. (2013): *Cloud Computing.*Tata McGraw Hill. 4. Jamsa, K. (2012):*Cloud Computing: SaaS, PaaS, IaaS, Virtualization, Business Models, Mobile, Security and More*. Jones & Bartlett Learning. Sosinsky, B. (2010): *Cloud Computing Bible.*Wiley-India, 2010.

M.Sc. (Data Science): IV Semester		
Course Code: MDS 405	**Lab Work Based on MDS 401, 402, 403, 404**	**Credit: 4** **Max Marks: 60+40 =100**
Course objectives:	To learn real life/industry applications of theory	
Course outcomes:		