M.Sc. DATA ANALYTICS
SCHEME OF EXAMINATIONS

BHARATHIAR UNIVERSITY
COIMBATORE – 641 046

Department of Computer Applications

MASTER OF SCIENCE IN DATA ANALYTICS – M.Sc. Data Analytics *(CBCS)*

1. Eligibility for Admission
A pass in any Bachelor's degree of minimum 3 years duration with Mathematics or Statistics as any one of the subjects at Graduate level.

2. Duration
The programme shall be offered on a full-time basis. The programme will consist of three semesters of course work and laboratory work and the fourth semester consists of project.

3. Regulations
The general Regulations of the Bharathiar University Choice Based Credit System Programme are applicable to these programmes.

4. The Medium of Instruction and Examinations
The medium of instruction and Examinations shall be in English.

5. Submission of Record Notebooks for Practical Examinations & Project Viva-Voce.
Candidates taking the Practical Examinations should submit bonafide Record Note Books prescribed for the Examinations. Otherwise the candidates will not be permitted to take the Practical Examinations. Candidates taking the Project Viva Examination should submit Project Report prescribed for the Examinations. Otherwise the candidates will not be permitted to take the Project Viva-voce Examination.

6. Ranking
A candidate who qualifies for the PG Degree Course passing all the Examinations in the first attempt, within the minimum period prescribed for the Course of Study from the date of admission to the Course and secures 1st or 2nd Class shall be eligible for ranking and such ranking will be confined to 10% of the total number of candidates qualified in that particular subject to a maximum of 10 ranks.

7. Revision of Regulations and Curriculum
The above Regulation and Scheme of Examinations will be in vogue without any change for a minimum period of three years from the date of approval of the Regulations. The University may revise /amend/ change the Regulations and Scheme of Examinations, if found necessary.

BHARATHIAR UNIVERSITY, COIMBATORE 641046.
Master of Science in Data Analytics (Univ.Dept.)

(Effective from the academic Year 2016-2017)
Scheme of Examinations

Core/ Elective/ Supportive / Project	Suggested Code	Sem	Title of the Paper	L	P	Internal	External	Credits	Marks
Core	16CSEBC01	I	Principles of Data Science	4	0	25	75	4	100
Core	16CSEBC02	I	Mathematical Foundations for Data Processing	4	0	25	75	4	100
Core	16CSEBC03	I	Linux Operating System (3 +1)	3	2	25	75	4	100
Core	16CSEBC04	I	Python Programming (2+2)	2	4	25	75	4	100
Core	16CSEBC05	I	Hadoop and MapReduce	4	0	25	75	4	100
Elective	16CSEBE01	I	Data Analysis using Pig and Hive (2+2)	2	4	25	75	4	100
Supportive	16CSEBGXX	I	General Supportive			12	38	2	50
Core	16CSEBC06	II	Data Warehousing and Data Mining	4	0	25	75	4	100
Core	16CSEBC07	II	Applied Statistics	4	0	25	75	4	100
Core	16CSEBC08	II	Exploratory and Descriptive Data Analytics (2+2)	2	4	25	75	4	100
Core	16CSEBC09	II	R Programming (2+2)	2	4	25	75	4	100
Core	16CSEBC10	II	NoSQL - MongoDB (2+2)	2	4	25	75	4	100
Elective	16CSEBEXX	II	Elective-I			25	75	4	100
Supportive	16CSEBGXX	II	General Supportive			12	38	2	50
Core	16CSEBC11	III	Data Processing for Data Analysis	4	0	25	75	4	100
Core	16CSEBC12	III	Machine Learning for Data Analytics (2 +2)	2	4	25	75	4	100
Core	16CSEBC13	III	Predictive and Inferential Analytics (2+2)	2	4	25	75	4	100
Core	16CSEBC14	III	Big Data Analytics in Cloud	4	0	25	75	4	100
Core	16CSEBC15	III	IoT and Big Data	4	0	25	75	4	100
Elective	16CSEBEXX	III	Elective-II			25	75	4	100
Mini Project	16CSEBC16	III	Mini Project			12	38	2	50
Project	16CSEBC17	IV	Project			75	225	12	300
								90	2250

Core/ Elective/ Supportive/ Project	Suggested Code	Sem	Title of the Paper	L	P	Internal	External	Credits	Marks
Elective I - Social Media Analytics									
Elective	16CSEBE02		Sentiment Analysis	4	0	25	75	4	100
Elective	16CSEBE03		Social Media Mining	4	0	25	75	4	100
Elective	16CSEBE03		Text Analytics using STORM	4	0	25	75	4	100
Elective	16CSEBE04		NoSQL Neo4j **(2+2)**	2	4	25	75	4	100
Elective II - Advanced Data Analytics									
Elective	16CSEBE05		Data Analytics using KNIME **(2+2)**	2	4	25	75	4	100
Elective	16CSEBE06		Data Visualization (**2+2)**	2	4	25	75	4	100
Elective	16CSEBE07		Streaming Analytics with SPARK **2+2)**	2	4	25	75	4	100
Elective	16CSEBE08		R-Hadoop and Hive	4	0	25	75	4	100

Course Title: PRINCIPLES OF DATA SCIENCE

Course Number: **Number of Credits: 4**

Subject Description
Goal
Objectives

Contents
UNIT I:
Data Evolution: Data Development Time Line – ICT Advancement-a Perspective – Data Growth-a Perspective – IT Components-Business Process – Landscape-Data to Data Science –

UNIT II:
Understanding data: Introduction – Type of Data: Numeric – Categorical – Graphical – High Dimensional Data — Data Classification – Hot Data – Cold Data – Warm Data – Thick Data – Thin Data - Classification of digital Data: Structured, Semi-Structured and Un-Structured. Sources of Data: Time Series – Transactional Data – Biological Data – Spatial Data – Social Network Data – Data Evolution – Data Sources

UNIT III:
Data Science: Data Science-A Discipline – Data Science vs Statistics, Data Science vs Mathematics, Data Science vs Programming Language, Data Science vs Database, Data Science vs Machine Learning. Data Analytics - – Relation: Data Science, Analytics, Big Data Analytics. Data Science Components: Data Engineering, Data Analytics-Methods and Algorithm, Data Visualization

UNIT IV:
Big Data: Digital Data-an Imprint: Evolution of Big Data – What is Big Data – Sources of Big Data. Characteristics of Big Data 6Vs – Big Data Myths - Data Discovery-Traditional Approach, Big Data Technology: Big Data Technology Process – Big Data Exploration - Data Augmentation – Operational Analysis – 360 View of Customers – Security and Intelligence

UNIT V:
Big Data Usecases –Big Data Technology Potentials – Limitations of Big Data and Challenges- Big Data Roles Data Scientist , Data Architect, Data Analyst – Skills – Case Study : Big Data – Customer Insights – Behavioral Analysis – Big Data Applications - Marketing – Retails – Insurance – Risk and Security – Health care

Reference Books

1. V. Bhuvaneswari, T. Devi, “Big Data Analytics: A Practitioner’s Approach” 2016.
2. Han Hu, Yonggang Wen, Tat-Seng, Chua, Xuelong Li,“Toward Scalable Systems for Big Data Analytics: A Technology Tutorial”, IEEE, 2014.

Course Title: MATHEMATICAL FOUNDATION FOR DATA PROCESSING

Course Number: **Number of Credits: 4**

Subject Description
Goal
Objectives

Contents

UNIT I:

Introduction to the concept of analytic function: Limits and continuity–Analytic functions –Polynomials–Rational functions –Conformality: Arcs and closed curves –Analytic functions in regions –Conformal mapping –Length and area –Linear transformations: The linear group–The cross ratio –Elementary conformal mappings: Elementary Riemann surfaces.

UNIT II:

Fundamental theorems: Line integrals rectifiable arcs –Line integrals as functions of arcs–Cauchy's theorem for a rectangle -Cauchy's theorem in a disk, Cauchy's integral formula: The index of a point with respect to a closed curve –The integral formula –Higher derivatives -Local properties of analytical functions: Removable singularities, Taylor's theorem –Zeros and poles –The local mapping –The maximum principle –The general form of Cauchy's theorem: Chains and cycles.

UNIT III:

Power series Expansions: Weierstrass theorem –The Taylor series –The Laurent series–Partial fractions and factorization: Partial fractions –Infinite products -Canonical products. The Riemann mapping theorem: Statement and proof –Boundary behavior –Use of the reflection principle –Analytic arcs –Conformal mapping of polygons: The behavior at an angle The Schwarz –Christoffel formula –Mapping on a rectangle

UNIT IV:

Linear equations with constant coefficients: The second order homogeneous equations –Initial value problems –Linear dependence and independence -A formula for the Wronskian –The non-homogeneous equation of order two. Homogeneous and non–homogeneous equations of order n –Initial value problems –Annihilator method to solve a non–homogeneous equation –Algebra of constant coefficient operators.

UNIT V:

Linear equations with variable coefficients: initial value problems for the homogeneous equation-Solutions of the homogeneous equation –The Wronskian and linear independence –Reduction of the order of a homogeneous equation -Homogeneous equation with analytic coefficients –The Legendre equation. Linear equation n with regular singular points: Euler equation -Second order equations with regular singular points –Exceptional cases –Bessel equation.

Reference Books:

1. E.A. Coddington, "An Introduction to Ordinary Differential Equations" Prentice Hall of India Ltd., New Delhi. 1961
2. L.V. Ahlfors, "Complex Analysis" Third Edition, McGraw-Hill, New York. 1966.

Course Title: LINUX OPERATING SYSTEM

Course Number: **Number of Credits: 4**

Subject Description
Goal
Objectives
Contents

UNIT I

Evolution of Operating system - System Calls, System Programs, Processes-Process Concept, Process Scheduling, Operations on Processes, Interprocess Communication; Threads-. Process Synchronization - Critical Section Problem, Mutex Locks, Semophores, Monitors; CPU Scheduling and Deadlocks.

UNIT II

Main Memory-Contiguous Memory Allocation, Segmentation, Paging, 32 and 64 bit architecture Examples; Virtual Memory- Demand Paging, Page Replacement, Allocation, Thrashing; Allocating Kernel Memory, OS Examples. Mass Storage Structure- Overview, Disk Scheduling and Management; File System Storage-File Concepts, Directory and File and Disk Structure, Sharing and Protection

Unit III

Linux System- Basic Concepts; Linux Terminology – Community – Distributions - Linux File system Basics – Boot Process- Distribution Installation – Documentation – Gnu – Help – System Administration-Requirements for Linux System Administrator, Setting up a LINUX Multifunction Server, Domain Name System, Setting Up Local Network Services; Virtualization- Basic Concepts, Setting Up Xen,VMware on Linux Host and Adding Guest OS.

Unit IV

Linux command line – Basic operations – searching – Working with files - File operations – architecture – compressing files – File permissions - Transferring Files - Text Manipulation – cat – echo sed- awk – grep – Bash shell scripting – String manipulation – Boolean expression – case statement – looping

UNIT V

User Accounts – Environment variables – command aliases –Linux Text editors – Vi – Gedit – nano – Linux Security – Usage of root account – using sudo - limiting hardware access – working with passwords – Securing Boot Process and Hardware resources - Case study – Installation of any one Linux Distribution

Reference Books

1. Abraham Silberschatz, Peter Baer Galvin and Greg Gagne, “Operating System Concepts”, 9th Edition, John Wiley and Sons Inc., 2012.
2. William Stallings, “Operating Systems – Internals and Design Principles”, 7th Edition, Prentice Hall, 2011.
3. D M Dhamdhere, “Operating Systems: A Concept-Based Approach”, Second Edition, Tata McGraw-Hill Education, 2007.
4. http://nptel.ac.in/.
5. Linux for Beginners: An Introduction to the Linux Operating System and Command Line, Jason Cannon,

Course Title: PYTHON PROGRAMMING

Course Number: **Number of Credits: 4**

Subject Description
Goal
Objectives

Contents

UNIT I
Introduction to Python: Python Introduction - History of Python - Python features, Python interpreter, Overview of programming in Pytho - Python built in types, Arithmetic - Program input and output, Variables and assignment. - Advanced data types :Python Strings and string manipulation Assigning values in strings, String special operators, String formatting operators, Triple Quotes, Raw String, Unicode String, Build-in-String methods - Python List: Introduction - Accessing values in list, List manipulations, List Operations, Indexing, slicing & matrices - Python Dictionary -Introduction, Accessing values, Properties, Functions in Dictionary - Python Tuples: Introduction, Operation, Accessing, Function and methods in tuples and Data Type Conversion.

UNIT II
Python - Basic Operators: Arithmetic Operators, Comparison Operators, Logical (or Relational) Operators, Assignment Operators, Conditional (or ternary) Operators - Conditional Statement: Branching (if, else-if, nested),Looping: while statement, for statements, Control Statements: break, continue and pass Statements - Functions: Defining a function, Calling a function, Types of functions, Function Arguments Anonymous functions - Global and local variables. - Modules: Importing module, Math module Random module - Packages, Composition -Exception Handling

UNIT III
OOPs concept: Class and object – Attributes – Inheritance - Overloading, Overriding - Data hiding. Regular expressions - Match function, Search function, Modifiers, Patterns- Files: reading and writing files, methods of file objects - Standard library functions - dates and times

UNIT IV
GUI Introductions: Introduction to GUI Programming, Tkinter programming, Tkinter widgets - Database: Python database application programmers interface (DB-API), connection and cursor objects - Type objects and constructors - python database adapters - Visualization: Bar chart, Polar plot, Pie Charts, Histograms, Contour Plot, Heat Map.

UNIT V
Networking: Socket, Socket Module and methods - Client and server Internet modules – Multithreading - Web Services and XML. JSON and the REST Architecture - Web Programming: Creating simple web clients - Introduction to CGI, CGI module, building CGI applications - python web application frameworks: django.

REFERENCES

1. Core Python Programming by Wesley J. Chun, 2nd Edition ,Pearson Education
2. An Introduction to Python by Guido Van Russom, Fred L.Drake, Network TheoryLimited.
3. Beginning Python: From Novice To Professional By Magnus Lie Hetland, Second Edition.
4. Programming in Python 3 by Mark Summerfield, Pearson Education
5. Online version of An Introduction To Python
6. http://www.network-theory.co.uk/docs/pytut
7. http://docs.python.org/tutorial/
8. www.spoken-tutorial.org

Course Title: HADOOP AND MAPREDUCE

Course Number:

Number of Credits: 4

Subject Description
Goal
Objectives

Contents

UNIT I

Introduction To Big Data: Introduction to BigData Platform–Traits of Big data-Challenges of Conventional Systems- - Web Data–Evolution Of Analytic Scalability-Analytic Processes and Tools-Analysis vs Reporting-Modern Data Analytic Tools – Data Processing Models – Limitation of Conventional Data Processing Approaches

UNIT II

Hadoop: Basic Concepts-An Overview of Hadoop-The HadoopDistributed File System-Anatomy of a Hadoop Cluster-Hadoop Ecosystem Components. **HDFS**: Hadoop distributed File System-HDFS Design and Architecture-HDFS Concepts – Interacting HDFS using commandline -Interacting HDFS using JavaAPIs – Dataflow – Blocks – Replica-Hadoop Processes-Name node-Secondary name node-Job tracker-Task tracker-Data node – Hadoop YARN – SPARK – STORM

UNIT III

HBASE: What is HBase?-HBase Architecture-HBase API-Managing large data sets with HBase – HBase in Hadoop applications -**Map Reduce:** Developing Map Reduce Application - Phases in Map Reduce Framework - Map Reduce Input and Output Formats - Advanced Concepts - Sample Applications – Combiner – Joining datasets in Mapreduce jobs – Map - side join – Reduce - Side join - Map reduce – customization

UNIT IV

Map Reduce Program: Introduction to Writing a MapReduce Program - The MapReduce Flow - Examining a Sample MapReduce Program- Basic MapReduce API Concepts - The Driver Code - The Mapper - The Reducer - Hadoop's Streaming API - Using Eclipse for Rapid Development –The New MapReduce API

UNIT V

Common Map Reduce Algorithms: Sorting and Searching – Indexing - Machine Learning With Mahout - Term Frequency –Inverse Document Frequency - Word Co-Occurrence. Hadoop Map Reduce Programming Languages: HIVE–HIVE-Map Reduce Programs through HIVE-HIVE Commands-Loading, -Sample programs in HIVE – PIG – Basics-Installation and Configurations-Command

Reference Book:

1. Seema Acharya, Subhashni Chellappan, "Big Data Analytics", Wiley, 2015.
2. Han Hu, Yonggang Wen, Tat-Seng, Chua, Xuelong Li,"Toward Scalable Systems for Big Data Analytics: A Technology Tutorial", IEEE, 2014

Course Title: DATA ANALYSIS USING PIG AND HIVE

Course Number: **Number of Credits: 4**

Subject Description
Goal
Objectives

Contents
UNIT I

Introduction to Big Data – Distributed file system –, Map Reduce Algorithm- Hadoop Storage [HDFS], Common Hadoop Shell commands - Anatomy of File Write and Read, NameNode, Secondary NameNode, and DataNode - Hadoop Configuration – Pig Configuration – Hive Configuration

UNIT II
Pig Introduction : Overview of Pig - Pig Architecture - Pig Execution modes, Pig Grunt shell and Shell commands. Pig Latin Basis: Data model, Data Types, Operator - Pig Latin Commands - Load & Store, Diagnostic Operators, Grouping, Cogroup, Joining, Filtering, Sorting, Splitting - Built-In Functions, User define functions.

Unit III
Pig Execution Modes – Batch Mode – Embedded Mode – Pig Execution in Batch Mode – Embedding Pig in Python – Use cases - Map Reduce programs with Pig – Pig Vs SQL

UNIT IV
Introduction of Hive - Hive Features - Hive architecture -Hive Meta store - Hive data types - Hive Tables - Table types - Creating database , Altering database, Create table, alter table, Drop table, - Built-In Functions - Built-In Operators, User defined functions, - View – Pig Vs Hive

UNIT V
HiveQL–Introduction to HiveQL, HiveQL Select, HiveQL – MapReduce using HiveQL OrderBy, Group By Joins, LIMIT, Distribute By , Cluster By - Sorting And Aggregation – Partitioning – Static – Dynamic – Index Creation - Bucketing – Analysis of MapReduce execution – Hive Optimization – Setting Hiivng Parameters. – Usecase : MapReduce using Hive QL – HiveQL Vs SQL

References

1. Boris Lublinsky Kevin T. Smith Alexey Yakubovich ,PROFESSIONAL Hadoop® Solutions , Wiley, ISBN: 9788126551071, 2015.
2. Chris Eaton, Dirk deroos et al. , “Understanding Big data ”, McGraw Hill, 2012.
3. Tom White ,“Hadoop: The Definitive Guide”, O'Reilly Media 3rd Edition,May6, 2012
4. Chuck Lam , “Hadoop in Action” ,Manning Publications; 1st Edition ,December, 2010
5. Donald Miner, Adam Shook, “MapReduce Design Patterns”, O'Reilly Media ,November 22, 2012
6. Edward Capriolo ,Dean Wampler ,Jason Rutherglen, “Programming Hive”, O'Reilly Media; 1 edition , October, 2012
7. Alan Gates , “Programming Pig”, O'Reilly Media; 1st Edition ,October, 2011

Course Title: DATA WAREHOUSING AND MINING

Course Number: **Number of Credits: 4**

Subject Description
Goal
Objectives

Contents

UNIT I

Data warehousing: Introduction - Definition - Multidimensional data model - OLAP operations - Warehouse schema - Data warehousing architecture - Warehouse Schema - Warehouse server - Meta data - OLAP Engine - Data warehouse backend process - Data Warehouse Technology - Warehousing Software - Cloud data warehousing - Other features. Data Warehousing Case Study: Government, Tourism and Industry

UNIT II

Data mining: Introduction – Data as a Subject - Definitions- KDD vs. Data mining- DM techniques- Current Trends in Data Mining. Association Rules: Concepts- Methods to discover Association rules- A priori algorithm – Partition algorithm- Pioneer search algorithm –Dynamic Item set Counting algorithm- FP-tree growth algorithm-Incremental algorithm-Border algorithm-Generalized association rule. Analysis of association rule using orange.

UNIT III

Clustering techniques: Data Attribute Types – Data Similarity and Dissimilarity - Clustering paradigms – Partition algorithm-K- Medeoid algorithms – CLARA- CLARANS –Hierarchical DBSCAN-BIRCH- CURE-Categorical clustering algorithms-STIRR-ROCK-CACTUS-Other techniques: Implementation of Clustering techniques using orange tool.

UNIT IV

Classification Technique: Introduction – Decision Trees: Tree Construction Principle – Attributte Selection measure – Tree Pruning - Decision Tree construction Algorithm – CART – ID3 - Rainforest - CLOUDS - BOAT, Pruning Technique – Model Evaluation –Cross Validation – Bootstrap – Holdout – Classifier Performance- Boosting – AdaBoost - Bagging

UNIT V

Web mining: Basic concepts – Web content mining – Web structure mining – Web usage mining – text mining – text clustering - Temporal and Sequential Data mining: Temporal Association rules – Sequence Mining – The GSP algorithm – SPADE – SPIRIT – WUM – Spatial mining – Spatial mining tasks – Spatial clustering – Spatial trends. Various tools and techniques for implementing (Weka, Rapidminer and Matlab)

Reference Books

1. Jiawei Han, Micheline Kamber, “Data Mining Concepts and Techniques”, Morgan Kaufmann Publishers, 2012
2. Arun K Pujari, “Data Mining Techniques”, Universities Press. 2012

Course Title: APPLIED STATISTICS

Course Number: **Number of Credits: 4**

Subject Description
Goal
Objectives

Contents

UNIT I

Introduction: Meaning, Definition, Statistics as a data, Statistics as a Method. Importance, Functions, and Limitations of Statistics in Data Science. Finite and Infinite population, Hypothetical and existent population, census method, sample method, Random sampling (Non-Random sampling, simple Random Sampling, Restricted Random Sampling), Statistical Sampling, Systematic Sampling, Clustering Sampling, Judgment Sampling, Quota Sampling, Convenience or Churk Sampling, Statistical Errors, Absolute Error, Relative error, Reducing Sample Error, Test of Reliability Error.

UNIT II

Classification and Tabulation: Overview of Classification, Statistical Series, Types of Series, Frequency Distribution, Continuous or Grouped Frequency Distribution. Magnitude of Class intervals, Cumulative Frequency Distribution, Two Way Frequency Distribution. Measures of Central Tendency:Arithmetic Mean, Geometric Mean, Harmonic Mean, Median, Mode. Dispersion: Overview, Mean Deviation, Standard Deviation, Combined Standard Deviation.

UNIT III

Correlation: Overview, Types of Correlation, Karl Pearson's Coefficient Correlation, Correlation and Probable Error, Rank Coefficient Correlation. Regression: Overview, Correlation and Regression, Graphical Method, Algebraic Method, Regression Line, Regression Equation, Mathematical Properties, Standard Error of Estimate. Association of attributes: Introduction, Classification, Correlation and Association, Types of Association, Comparison of Observed and Expected Frequencies, Yule's Coefficient of Association, Yule's Coefficient of Colligation, Pearsons' Coefficient of Contingency Partial Association.

UNIT IV

Probability: Introduction, Mathematical Properties, Permutation, Combination, Trail, Sample Events, Sample Space, Mutually Exclusive Cases, Exhaustive Events, Independent Events, Dependent Events, Simple and Compound Events, Classical, Relative Frequency, Theory of Probability, Personalistic view of Probability, Addition and Multiplication Theorem, odds. Theoretical Distribution: Binominal Distribution, Obtaining Coefficient, Poison Distribution, Normal Distribution.

UNIT V

Sampling Theory and test of significance: Introduction Estimation, Hypothesis, Standard Error, Test of Significance for Attributes, Test of Significance for Large Samples. Test of Significance for Small Samples. Chie Square Test: Introduction, Assumption, Uses of X^2Test of Goodness of fit, X^2 Test of Independence, Yate's Correction, X^2 test of Homogeneity, Additive Property

Reference

1. R.S.N. Pillai, Bagavathi, "Statistics Theory and Practice, S.Chand& Company, 2013
2. Douglas C. Montgomery, George C. Runger., "Applied Statistics for Engineers", John Wiley & Sons. Inc, 2003

Course Title: EXPLORATORY AND DESCRIPTIVE DATA ANALYTICS

Course Number: **Number of Credits: 4**
Subject Description
Goal
Objectives

Contents

Unit I

Introduction – Data Analytics – EDA – Need for EDA – EDA – Objectives - Google Trend analysis – Explore trends - R Visualization – Packages – Lattice – ggplot2 – understanding plots – aesthetics - - statistical function - Histogram – Box Plot – Density Plot – Scatter Plots – Summarizing Data in R

Unit II

Variable Analysis – One variable – Understanding outliers through – histogram , boxplot, density plot – dataset – pseudo dataset of facebook Exploring two variables – Understanding Variables and relationships – scatter plots – correlations – condition means – Explore multivariate variables – Visualization of variables using aesthetics in R – Case study – Explore Diamond dataset for prize prediciton

Unit III

Data types – Categorical – Binary – ordinal – Nominal – Continuous – Discrete – Data Dimensions – Univariate – bivariate – multivariate – Numerical Measures – Central Tendency – Mean – Median – Mode - Understanding data using central tendency – plotting histogram – density plots and inference of plot - Variability Measure – Variance - Range - IQC - and Standard Deviation – Sum of squares – Squared Deviations – Absolute Deviations - Identify outlier using Inter Quartile Range – Visualization using boxplot

Unit IV

Data standardizing – Z Score – Negative Z Score – Continuous Distributions - Compute proportions – Relative Frequency histogram - Normalized Distribution using Ztable – Probability Distributions - Probability of mean – location of mean distribution - Sampling Distributions — Klout Sampling Distribution – Understanding Shape of Distribution – Standard Error - Standard Deviation of sampling distribution – Ratio of Sampling Distribution - Central Limit Theorem R – Mean of sample means Advanced Analytics

Unit V

Case Study – EDA analytics on dataset Movies – Social network using R – Prediction of Movie ratings – Descriptive Analytics on Movie Dataset

Reference Books

1. Hadley Wickham, ggplot2: Elegant graphics for data analysis, Springer (2009) http://www.springerlink.com.proxy.lib.umich.edu/content/978-0-387-98140-6/contents/
2. Phil Spector, Data Manipulation with R, Springer , (2008http://www.springer.com/statistics/computanional+statistics/book/978-0-387-74730-9
3. Leland Wilkinson, The Grammar of Graphics, Springer (2005),http://www.springerlink.com.proxy.lib.umich.edu/content/978-0-387-24544-7/contents/
4. Statistical Inference. Casella, G. and Berger, R. L. (1990). Wadsworth, Belmont, CA.

Course Title: R PROGRAMMING

Course Number: **Number of Credits: 4**

Subject Description
Goal
Objectives
Contents

UNIT I:

Introduction: What is R – Downloading and Installing R – Script Code – Graphical Facilities in R – Editors – Help Files and New groups – Packages - General Issues in R. **Getting Data into R:** First Step in R: Typing in Small Datasets – Concatenating Data with c Function – Combining Variables with the c, cbind, rbind Functions – Combining Data with the Vector Function – Combining Data Using Matrix – Combining Data with data frame - Function – Combining Data Using the List Function – Importing Data: Importing Excel Data – Accessing Data from other Statistical Packages – Accessing the Database.

UNIT III:

Accessing Variables and Managing Subsets of Data: Accessing Variables from Data Frame: The str Function – The Data Argument in Function – The $ sign – The Attach Function. Accessing Subsets of Data – Combining Two Datasets with a Common Identifier – Exporting Data – Recording Categorical Variables. Simple Functions **:** The Tapply Function – The Supply and Lapply Function – The Summary and Table Function.

UNIT III:

Importing Data – Csv, Excel, Table, Xml, Json , Databases Compare Vector – Matrices - Conditional – Control flow – Loops – A Function with Multiple Arguments - Cleaning Data : – Exploring raw data – Missing values - Zeros and NAs – Separating – Uniting Columns - String Manipulation – Filling Missing values – Packages – Dyplr – Statistical Functions – Comparison of Samples – same groups – different groups – Independent groups - Student T Test – Dependent Test – Independent Test

UNIT IV:

Basic Plotting Tools: Loops and Functions: Introduction to Loops – Loops: Importing Data – Making Scatter Plot and Adding Labels – Designing General Code – Saving the Graph – Construction the Loop, The Plot Function – Symbol, Colors and Sizes – Adding a Smoothing Line The Pie Chart – The Bar and Strip Chart – Box Plot – Cleveland Dotplots – Revisiting the Plot Function: More Options for Plot Function – Legends - – The Pair Plot – The Coplot – Combining Types of Plots. An Introduction to Packag**e:** High Level Lattice Function – Ggplot2

UNIT V:

Reporting – Data Prepration – Embedding R chunks – Labelling and resuing code chunks – Report Compiling – Configuring – R Packages – shiny - ggvis - Common R Mistakes: Problems Importing Data – Attach Misery – Non-Attach Mesery – The Log of Zero - Miscellaneous Errors – Mistakenly Saved R Workspace.

Reference:

1. Alain F. Zuur, Elena N. Ieno, Erik H.W.G. Meesters, “A Beginner’s Guide to R” Springer, 2009
2. Roger D. Peng, “R Programming for Data Science” Lean Publishing, 2014
3. R Data camp – Online Course Contents - https://campus.datacamp.com/courses/

Course Title: NoSQL MongoDB

Course Number: **Number of Credits: 4**

Subject Description
Goal
Objectives

Contents

Unit I

Big Databases- SQL – NoSQL Tradeoffs – CAP Theorem – Eventual Consistency - NoSQL – database types – MongoDB- Introduction - MongoDb – Need – MongoDB Vs RDBMS – MongoDB- Driver Installation – Configuration – Import and Export – MongoDB Server Configuration

Unit II

Data Extraction Fundamentals - Intro to Tabular Formats - Parsing CSV -Parsing XLS with XLRD- Parsing XML - Intro to JSONGetting Data into MongoDB - MongoDB- CURD – Database Creation – Update – Read –Delete Using mongoimport -Operators like $gt, $lt, $exists, $regex -Querying Arrays and using $in and $all Operators -Changing entries: $update, $set, $unset

Unit III

Data Analysis - Field Queries -Projection Queries- Limiting – Sorting - - Aggregation - Examples of Aggregation Framework -The Aggregation Pipeline -Aggregation Operators: $match, $project, $unwind, $group –

Unit IV

User Management – MongoDb Data Replication in Servers – Data Sharding – MongoDB Indexes – Create – Find – Drop – Backup – MongoDB – Relationships – Analyzing Queries – MongoDB Objectid

Unit V

Advanced MongoDB: MapReduce – MongoDB - Text Processing - Regular Expression – Case Studies – Text processing of large datasets, Map Reduce using MongoDB

References

1. MongoDB: The Definitive Guide, 2nd Edition , Powerful and Scalable Data Storage, By Kristina Chodorow, Publisher: O'Reilly Media
2. MongoDB Basics - **David Hows, Peter Membrey, Eelco Plugge, Publisher Apress - Ebook(free) https://it-ebooks.info/book/4527/**

Course Title: DATA PROCESSING FOR DATA ANALYSIS

Course Number: **Number of Credits: 4**

Subject Description
Goal
Objectives

Contents

Course Title: MACHINE LEARNING FOR DATA ANALYTICS

Course Number: **Number of Credits: 4**

Subject Description
Goal
Objectives

Contents

Course Title: PREDICTIVE AND INFERENTIAL ANALYTICS

Course Number: **Number of Credits: 4**

Subject Description
Goal
Objectives

Contents

Course Title: BIG DATA ANALYTICS IN CLOUD

Course Number: **Number of Credits: 4**

Subject Description
Goal
Objectives

Contents

Course Title: **IOT AND BIG DATA**

Course Number: **Number of Credits: 4**

Subject Description
Goal
Objectives

Contents

Elective I - Social Media Analytics

Course Title: **SENTIMENT ANALYSIS**

Course Number: **Number of Credits: 4**

Subject Description
Goal
Objectives

Contents

Course Title: **SOCIAL MEDIA MINING**

Course Number: **Number of Credits: 4**

Subject Description
Goal
Objectives

Contents

Course Title: **TEXT ANALYTICS USING STORM**

Course Number: **Number of Credits: 4**

Subject Description
Goal
Objectives

Contents

Course Title: NoSQL Neo4j

Course Number: **Number of Credits: 4**

Subject Description
Goal
Objectives

Contents

Elective II - Advanced Data Analytics

Course Title: DATA ANALYTICS USING KNIME

Course Number: **Number of Credits: 4**

Subject Description
Goal
Objectives

Contents

Course Title: DATA VISUALIZATION

Course Number: **Number of Credits: 4**

Subject Description
Goal
Objectives

Contents

Course Title: STREAMING ANALYTICS WITH SPARK

Course Number: **Number of Credits: 4**

Subject Description
Goal
Objectives
Contents

Course Title: R-HADOOP AND HIVE

Course Number: **Number of Credits: 4**
Subject Description
GoalObjectives
Contents